DOCUMENT RESUME

ED 292 811 TM 011 115

AUTHOR Corbett, H. Dickson; Wilson, Bruce L.
TITLE Study of Statewide Mandatory Minimum Competency Tests.
INSTITUTION Research for Better Schools, Inc., Philadelphia, Pa.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 30 Nov 87
NOTE 119p.
PUB TYPE Reports - Descriptive (141) -- Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Academic Achievement; Administrator Attitudes; Educational Change; Educational Testing; Elementary Secondary Education; *Minimum Competency Testing; School Districts; School Personnel; School Surveys; *State Legislation; State Programs; Student Evaluation; *Teacher Attitudes; Testing Programs

ABSTRACT

The reactions of local educators (teachers, principals, and central office personnel) to the initiation of statewide mandatory minimum competency tests and the effects of implementing these testing programs were studied. Variations among school districts within two states were considered. Items from a survey of individual educators were combined into scales measuring local system adjustments. Scales were used to study a "low stakes" state, where relatively minor consequences depended on student performance, and a "high stakes" state where graduation depended on passing the tests. The program had the greatest impact in the high stakes situation. For individual districts, a high stakes strategy had desirable consequences as long as the districts were not under too much pressure. District variations were also a result of the perceived political climate between the district and state department. Some positive results attending the state testing programs include better definition of curriculum, availability of additional information on students, and perception that students' skills increased. Data are summarized in charts in an appendix, and the survey questionnaire is also appended. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ERIC TM

ED292811

STUDY OF STATEWIDE MANDATORY MINIMUM COMPETENCY TESTS

H. DICKSON CORBETT

BRUCE L. WILSON

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ELLEN NEWCOMBE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Research for Better Schools
444 North Third Street
Philadelphia, PA 19123

November 30, 1987

TM 011 115

BEST COPY AVAILABLE

Study of Statewide Mandatory Minimum Competency Tests

H. Dickson Corbett
Research for Better Schools

Bruce L. Wilson
Research for Better Schools

November 30, 1987

The preparation of this paper was supported by funds from the U.S. Department of Education, Office of Educational Research and Improvement (OERI). The opinions expressed do not necessarily reflect the position or policy of OERI, and no official endorsement should be inferred. Appreciation is expressed to Robert Stake, Jane Hannaway, and John Connolly for insightful comments and to Arlene Ziviello for wordprocessing.

FINAL REPORT: RBS' STUDY OF STATEWIDE MANDATORY MINIMUM COMPETENCY TESTS

Close to sixty percent of the states in this country have mandated some form of standardized testing for local schools (Marshall, 1987). Nevertheless, debate continues about the consequences of implementing testing programs. The effects of assessment initiatives are not clear and have not been well informed by empirical research (Airasian, 1987; Rosenholtz, 1987; Stake, Bettridge, Metzer, & Switzer, 1987). Little is known about how the daily lives of public school staff and students are affected by statewide standardized testing; even less is known about how differences in state programs and school district characteristics magnify or minimize the effects.

This study belongs in the genre of research projects that examine assessment effects on local educational agencies (LEA)--that is, the study of the intended and unintended consequences of implementing assessment programs. The study had three purposes: (1) to gather local educator's reactions to the initiation of statewide, mandatory minimum competency testing in their respective states; (2) to compare the instructional, organizational, and cultural effects of implementing these testing programs on local school systems in two states; and (3) to explain district-to-district variations in effects within each state.

Findings related to the <u>first</u> purpose are presented in the form of a "Gallup Poll." Educators' responses to selected, individual items from a survey conducted in the two states are reported. The opinions are broken down by state and by position of the respondent (teacher, principal, central office) and are presented in five categories: local context, the tests, purposes, strategies, and impacts.

Individual survey items were combined into scales measuring various local system adjustments to facilitate the between-state comparisons that addresss the study's second purpose. The two states represented "low stakes" and "high stakes" situations (Madaus, in press). Relatively minor consequences attended student performance on State A's minimum competency tests in language and math. The purpose of both tests originally was to identify students needing additional classroom instruction who may not have been identified by other means used in the district. State B's "high stakes" strategy required students to pass reading, writing, math, and citizenship minimum competency tests in order to receive a high school diploma. The tests were being phased in as graduation requirements; at the time of the survey only the reading and math tests "counted." The tests had been preceded by several years of curriculum development intended to soften the blow of meeting the testing requirements. A substantial portion of the section of the report that addresses the high stakes/low stakes distinction will be devoted to the effects of events after the survey was conducted. In State A, for example, the stakes rose. The Chief State School Officer released the test scores by school prior to the 1987-88 school year and touted the test as an indicator of the schools' effectiveness or lack thereof. Study interviews conducted subsequent to this event revealed considerable concern on the part of school staff that the stakes were being raised even though the rankings were quickly withdrawn. State B had no similar dramatic event; instead its districts had to reconcile themselves to the inevitable day when all four tests would affect whether students graduated.

Implementation effects were not uniform across the school systems within each of the two states and the third purpose of the study was to explain these differences. Three sets of local adjustments to the testing program were the

focus of this analysis: curriculum and instruction, student impacts, and the quality of teachers' work lives. Explanations for differential impact in these three categories resided in other instructional and organizational adjustments the school systems made, demographic characteristics of the systems, testing program prcedures (especially in light of on-going testing in districts), and the LEA's relationship to the state education agency (SEA).

Before turning to the actual findings related to the study's three purposes, this report first reviews the scant but growing literature on assessing the effects of assessment, describes the conceptual framework that guided data collection and analysis, and details the research methods used.

Research on the Implementation of Minimum Competency Tests

"The crucial issues of testing are not technical. Issues of testing today are social, economic, and value-laden, involving the distribution and redistribution of resources and prerogatives" (Airasian, 1987:408). Research on minimum competency testing (MCT), defined in policy terms as "a device for conditioning student promotion or graduation on test achievement" (Darling-Hammond & Wise, 1985:318), has not yet caught up with this argument. Indeed, despite considerable state use of some form of competency testing as a policy tool (Marshall, 1987), most research continues to focus on technical issues related to the design of valid tests and the improvement of scores. Policy research is only beginning to address the implications of implementing such programs at the local level.

Madaus (in press) and Airasian (1987) make important distinctions concerning the use of minimum competency tests. Madaus suggests that standardized tests can be characterized as "high stakes" or "low stakes." High stakes tests are used for important decisions such as promotion or

graduation. For this reason, they have the ability to influence system behavior, even to direct it. In low stakes situations, no important sanctions follow test performance and thus the tests likely would have little effect on the system.

Airasian (1987) claims that standardized testing once served very general purposes; it was a tool to be used for both identifying areas where instruction needed improvement and gauging how well the educational system as a whole was functioning. More recently, according to Airasian, the success of these traditional uses of the tests have led to acceptance of a new purpose.

> This second use is most aptly termed state-mandated certification testing. In this approach, testing is not used to guide classroom instruction or to monitor educational policy. Rather, state-mandated certification testing has made testing and test results a crucial aspect of educational policy itself. (p. 403)

In other words, states have begun to use tests as the policy to try to spur improvements. These tests, of which MCT is one form, often have common characteristics: they are mandated for most students in selected grades; they eliminate local discretion by using one instrument to be administered and scored similarly across all systems; and they usually measure performance on a pass/fail basis. The consequences of such a testing policy are that test information becomes of interest to a wide population and not just a few professionals and concerned parents, local control over the curriculum may be eroded, and a tension is created between quality of education and equality of educational opportunity.

These two authors suggest that differences in the purposes of a statewide testing program and its concomitent consequences should have different impacts on their respective LEAs. The two states examined in this report were selected with that in mind. State A's approach was much more in line with

Airasian's (1987) assessment of the traditional use of standardized testing and contained low stakes for the system as whole, although remediation money was given to districts on the basis of how many student's fell below the cut-off point--a potential negative incentive for improving scores. State B, on the other hand, designed the test as a specific policy tool and tied student performance to high school graduation. Thus, the administrators, teachers, and students in this state faced a high stakes situation.

The available literature offers little guidance as to what precisely the differential impacts of such programs might be. Stake et al. (1987) provides an initial review of research on the effects of state assessment initiatives, examining the topic across six categories of effects: achievement standards; public attitude toward schools; the morale/motivation of those tested; the utility of test information for school administration; the reactions of teachers to standardized test results; and the curriculum. The review notes that few studies have been conducted to compare the local system consequences of statewide standardized (and/or minimum competency) testing programs.

The research that is available actually has focused more on general reactions to the tests than on the specific events, actions, and interpretations surrounding the implementation of testing programs at the state and local levels. Essentially two issues have been addressed in these studies: (1) the narrowing of the local curriculum and a concomitant loss of local input into curricular decisions and (2) alterations in the quality of life for teachers and students.

With regard to the first issue, what to some would be a "test-valid" curriculum (Popham, 1985) is to others a narrowed, and thus less satisfactory curriculum. In a study of 43 teachers from three districts, Darling-Hammond

and Wise (1985:319) reported that "some fear that, because the connection between the test and what it seeks to measure is tenuous, the means will substitute for the ends; the test will serve as the goal of instruction rather than as a measure of instruction or learning...Many teachers observe that, when they are pressured to 'teach to the test,' scores in the tested areas increase, but other types of learning suffer" (p.321). In addition, teachers "object to the standardization that results from the policies rather than the standards contained in the policies" (p.331). In the process, control over the curriculum gradually shifts to the administrators of the policy and/or the test makers. In the study reported above, teachers were only anticipating the consequences of a proposed MCT program. Shannon (1986) found that when reading test scores were actually used as the basis for a merit pay plan that teachers and administrators "standardized their definitions of reading as being equivalent to tested skills, constricted their methods of instruction to the use of a single procedure, and separated teachers intellectually and emotionally from their instruction" (p. 31). Essentially the researcher argued that the 34 teachers and five administrators in the district studied were "deskilled" (Apple, 1982)--their knowledge and expertise were ignored in favor of implementing uniform plans across the district.

With regard to the second issue, two studies have examined MCT and the quality of life in schools. Rosenholtz (1987) claims that teachers' commitment to teaching suffers under statewide MCT. Based on empirical research in a single state, the author says that the ways teachers view the task of teaching precludes the use of MCT as a valid measure of effectiveness. Thus, the onset of MCT exacerbates conflict between different views of teaching, and teacher commitment decreases as a result. With students, Serow and Davies (1982) found that minority students need additional remediation to

achieve a passing score on one state's MCT, particularly in reading. However, the remediation program was designed such that all failing students tended to be treated equally. The authors hint that such equal treatment may not correspond to equal opportunity.

While the amount of research on this topic is quite limited, the issues raised and the findings presented indicate the need to pursue the effects of both the assessment process as well as the results of particular assessment instruments, especially since the number of statewide testing programs in operation is so large (Marshall, 1987).

These programs have been and are being designed in the absence of systematic information about their intended and unintended consequences even though educators have been facing these statewide programs for several years. This literature review suggests a need for research that taps educators' reactions to the testing programs, documents differences in the effects of testing programs having different purposes, and examines the issue of how different districts are variously affected.

Conceptual Framework

The effects of introducing and operating a mandatory statewide testing program are expected to require adjustments in the instructional program, a district's organization, and a district's culture. An underlying assumption of this study is that the mandatory testing programs had far-reaching ramifications for the technology, structure, and values in place in school systems depending upon what was at stake. Potential instructional adjustments include: (1) strategies devised by a district to improve test scores, and (2) modification of curriculum and instruction to insure a better match with the content on the test. Organizational adjustments might be: (3) the extent to

which the test becomes a key indicator of system performance, and (4) the amount of information sharing the system engages in to inform staff about the tests. The cultural category of system adjustments includes impacts on: (5) student life, and (6) the quality of teacher worklife.

Whether or not adjustments in instruction, organization, and culture actually occur is dependent on at least two aspects of a system's operating environment. Summarized in Figure 1, these aspects are: (1) selected features of school district context, and (2) characteristics of the state testing program. Both constitute the major foci for analysis in this report.

With respect to school district context, years of research on educational change point to an inescapable conclusion: some programs work some times in some places, and it is mostly the time and the place that explain the fate of a program (Berman, 1981; Corbett, Dawson, & Firestone, 1984). Both Elmore (1980) and Berman (1981) argue that policy implementation can only be understood in terms of the context of the "target's" setting; policy makers' intentions become diffused and redirected as they pass through the prism of local politics, organization and culture. Thus, changes in the test scores over time are assumed to be the product of the complex interaction among system demographic and internal contextual characteristics, its relationship with the external environment--particularly the SEA, and the kinds of adjustments the system makes to implement the tests.

Features of the state testing program also will influence the type and magnitude of system adjustments that are made. The two programs to be examined in this study were mandatory but they varied in five important ways:

- the grade levels tested
- the type of standard used to determine minimum competence
- the local consequences for failure

Figure 1

CONCEPTUAL FRAMEWORK

SYSTEM ENVIRONMENT	SYSTEM ADJUSTMENTS	SYSTEM EFFECTIVENESS
1. District Context • internal contextual and demographic characteristics • district-SEA relationship	1. Instructional • strategies • curriculum & instruction	1. Student Focus • test scores • dropout rate • attendance • post-school plans
2. State Testing Program • high or low stakes - grade level focus - standards - consequences - assistance - use of results	2. Organizational • information flow • benchmark 3. Cultural • quality of work life • quality of student life	2. Teacher Focus • job satisfaction

- the kinds and amount of state assistance available
- what data comparisons were made public.

The essential difference was that the program in State B made graduation from high school dependent upon a student's passing writing, citizenship, reading, and math tests. In State A, the test was intended formally to serve the purposes Airasian (1987) identified as the traditional uses of standardized tests--namely, as a tool for fine-tuning classroom instruction to meet certain student needs and as a general overall indicator of educational performance in the state. Given the definition of MCT provided earlier, the comparison made in this study is between a more traditionally conceived, low-stakes testing program (State A) and a high-stakes MCT program having consequences for graduation (State B). State A's program was a recent legislative response to the educational reform movement spawned by the commission reports in the early 1980's whereas State B's testing program was preceded by a thorough curriculum improvement initiative specifically intended to boost local systems' programs so that the actual test, when implemented, would not be burdensome.

According to Madaus (in press) "high-stakes" programs are used for important decisions and thus have the power to modify decisions; "low-stakes" programs are not obligatory nor even generally anticipated to be central to decision-making, and test performance usually does not stimulate significant rewards or sanctions. The researchers selected the two states for study to accentuate the high-stakes/low-stakes distinction.

There are several reasons why higher stakes situations can be expected to have greater local impacts. First, mandatory tests are likely to force adjustments in a system by creating expectations for what the outcomes of schooling should be. According to Mintzberg (1983), stipulating outcomes is one means used widely in organizations to affect operations. Some standard--

no matter how narrowly defined--is to be met, regardless of what else staff members may want to accomplish. In situations where the standard is easily attained, its importance as a criterion of success may remain no more preeminent than any of a myriad of indicators. However, in situations where the standard is less readily met, its importance looms larger and perhaps more directly defines instruction and organization.

Second, one of schools' primary tasks is to move students smoothly through a series of grades to graduation (Schlechty, 1976). Staff size, the number of classrooms needed, and the availability of sufficient materials are all predicated in most communities on the assumption that essentially all first graders will become second graders and that most seniors will graduate on time. A few exceptions cause no problems, but testing programs change the assumptions by inserting an unpredictable checkpoint for determining progress for all students, based on something other than student age, credits obtained, or time spent in school. Obviously, some checkpoints are more formidable than others as in the case where successful completion of the test determines whether or not students graduate. But even relatively innocuous checkpoints may force some remediation and thereby affect subsequent progress.

Third, establishing a standard all students must meet as a visible indicator of effectiveness runs counter to the ethos of many educators (Rosenholtz, 1987). In spite of enormous standardization, a tone of individualism permeates American education (Lortie, 1975). Teachers are allowed considerable autonomy in determining what and how to teach, and they expect to handle their classrooms on their own. Moreover, many of the illustrations of successful schools that dot the literature rely almost solely on anecdotes about children who, despite special problems, manage to achieve in the classroom or life, reinforcing the individualistic notion that each

child has to be dealt with on his or her own terms. Testing programs challenge this ethos. Test items highlight critical content to cover, test administration dates determine the deadline for teaching the content, item formats affect how the information will be accessed, and the standards add a quality of sameness to what students should achieve. The tests, therefore, have major affects on school culture. Wilson (1971) defines culture as "definitions of what is and what ought to be . . ." Deal (1985) describes it as "the way we do things around here." Testing programs are likely to require serious examination of definitions of what being a student, teacher, or administrator is and should be. The literature on educational change is replete--although this is not always recognized--with descriptions of the clash between values implicit in an innovation and the values implicit in the way those expected to innovate were accustomed to behaving (Sarason, 1971; Gordon, 1984; Rossman, Corbett, & Firestone, forthcoming).

The conceptual framework provides a way to identify what system adjustments occurred and why they happened. It also points to an additional important question: Have the changes made the district more effective? Narrowly conceived, this question merely suggests an examination of a district's success in helping students meet the standards set by the test. However, it is becoming more and more clear that definitions of effectiveness and the extent to which they are shared are context dependent (Rossman, Corbett, & Firestone, 1985). Effectiveness, thus, may be defined more by how well a system prevents dropouts, improves attendance, stimulates student enthusiasm for learning, or addresses student differences than by doing better on a test. A two-year study is not an appropriate vehicle for answering this question. While the study can tap perceptions of a district's reach for improvement, the major focus of this report will be on explaining system

adjustments, not the ultimate effectiveness of the testing programs. The conceptual framework will be revisited at the end of this report and revised to reflect the study's findings.

Study Design

The conceptual framework simplifies a very complex situation. Introducing and operating a mandatory statewide testing program involves a wide range of potential challenges to a district. Many of these can be deduced from a conceptual framework such as the one above. However, using an inductive approach in which the research can take advantage of unexpected developments can be equally valuable (Miles & Huberman, 1984). For this reason, the study was designed to include both open-ended qualitative fieldwork and structured survey questionnaires. The study was conducted in three phases. First, a preliminary round of qualitative fieldwork was performed wherein researchers visited each school district for several days to talk to a wide variety of staff members. Second, the results from the interviews were used to design a questionnaire for broad use throughout the states studied. Third, the survey results were used to structure a final round of feedback and interviews in the sites originally visited.

Phase One: Fieldwork in 12 sites

The study was conducted in two states, chosen to highlight the high-stakes/low-stakes distinction. Six sites in each of the two states were visited. Site selection was made on the basis of district size and type of community served, primarily because these characteristics were assumed to determine the kind of staff resource demands implementing the test would make. Equally important was the willingness of the district to participate because

the purpose of this phase was to explore issues in depth, not to generalize strictly to a larger population. Selection was carried out with the input and assistance of key SEA staff members in each state.

Six experienced field researchers conducted the site visits. One researcher spent two or three days in each site (in late April and early May, 1986), depending on district size. The first day was spent in the central office, interviewing the superintendent (if available), the person(s) responsible for handling the testing program, and others on the district's staff. Also, pertinent documents were examined where available. School interviews were conducted with administrators, guidance counselors, teachers, and students. When all appropriate schools in a district could not be visited, selection was made in conjunction with site personnel. Sampling the variety of schools in a district was the foremost criterion. In all, over 250 local educators and students participated in the interviews. After preliminary data analysis occurred, a research brief was sent to each site in late June to share themes that emerged from the interviews.

Interview Questions. Field researchers operated from interview guides with broad categories of questions (see Figure 2). Specific phrasing of questions and the particular probes used were determined by the researcher on site. In training sessions conducted prior to the site visit, researchers had an opportunity to generate and discuss potential questions and follow-up probes, but fieldwork of this type demands that the researcher have considerable flexibility in determining who to talk to, what to ask, and when to ask it. The major information categories were drawn from the categories in the conceptual framework in Figure 1. Responses to questions were handwritten by the researcher.

Figure 2

SAMPLE INTERVIEW GUIDE

Information Categories	Potential Sources
1. Local Testing Program -The planning sequence -Staff involved -Direct contacts with SEA -Implementation sequence -Revisions in the subsystem so far -Arrangements for administration of the test	-Central office staff -Building level person responsible for test administration
2. State Testing Program -Grades tested -Standards -Local consequences -Comparisons -Assistance	-Documents -SEA background visits -Same people as #1 above
3. Internal District Context (both prior to program and effects) -Teacher autonomy in instructional decisions -Definitions of effective teaching -Pacing of instruction -Model student who is target of instruction -Supervision/evaluation patterns -Central office/building administration interest in instruction -Curriculum revision patterns -Views on role of standardized testing -Daily schedules	-Central office staff -Building administrators -Teachers -Students
4. Environmental Context -Relationship with SEA -Illustrative incidents with SEA -Views on role of SEA -Relationship with community -Illustrative incidents with community -Views on role of community involvement -Media attention to tests	-Central office staff -Selected building staff
5. Indicators of Effectiveness -Test scores -Alternative criteria generated by respondents	-Central office staff -Building administrators -Teachers -Students

Different information was collected from different staff members depending on their positions and responsibilities. The goal was to obtain data on each category from multiple sources but not necessarily from every source. Figure 2 contains a sample interview guide plus a listing of potential sources of information for each category.

Data Management. A multiple-case, multiple-researcher, open-ended interview study places a heavy burden on the data management system. A systematic way of determining data gaps, locating overlooked sources, making data accessible to other researchers, and being able to retrieve bits of data was imperative. To accomplish this, resources were allocated more to developing data summaries than to making handwritten field notes presentable or typing transcripts from tape recordings. When researchers returned from a site visit, they completed a series of data summary charts: (1) a summary of information sources and the categories for which each source supplied information; (2) a description of source-identified effects coupled with the researcher's designation of which and how many staff members listed each effect; (3) a summary of data on the district's instructional, organizational, and cultural contexts as well as its relationship with the surrounding community and the SEA; and (4) a listing of residual incidents and data worthy of note that did not fit cleanly in the structured charts.

These data summary charts were made available to the authors who did the cross-site analysis and they were the stimulus for determining whether additional information needed to be gathered from particular sites. Actual data summary charts are shown in Appendix A.

Data Analysis. The analysis activities consisted of reviewing the data summary charts to identify implementation themes that cut across the 12 sites. The specific goal of the analysis was to develop items for the questionnaire to be used in the second phase of the study.

Eight themes emerged from the researchers' extensive review of the data summary charts. These were:

- Degree of acceptance of the general idea of mandatory testing. Few staff quarreled vehemently with the appropriateness of a statewide test. "We need something like this" was a frequent refrain.
- At the same time, the tests' information was viewed as generally redundant in most districts, especially the suburban ones. That is, the MCT offered little information that the system did not already have available.
- "Teaching to the test" was a major concern and acknowledged as the most expedient means of trying to improve test scores. Perceptions about the "propriety" of the practice varied. Probably most heard was: "I don't believe in it but we have to do it to get scores up" another comment was "The tests cover important material so we should be teaching to the tests anyway". "I don't believe in it and won't do it" was also voiced.
- Staff members from systems and/or schools that did well on the tests were less unhappy about the program. Essentially they were pleased that the test scores gave the public confirmation of the good job they knew they were already doing.

- Related to this, in some systems that were doing well on the tests staff members also reported students had always done well on a variety of student performance measures. The test was not a strong stimulus for improvement. Instead, the test scores were the results of efforts undertaken before the tests were mandated.
- Socio-economic status of the community and community attitudes toward education were generally viewed as being a major determinant of test results.
- Dissatisfaction with state aspects of the program included: (1) test results were not provided in a format useful for classroom level decision-making; (2) test results were not provided at a time when much use or sense could be made of them, e.g., too late in the year, too early for a school to have made a difference, or too late for a school to make a difference; (3) the level of and/or specific skills covered in the tests were inappropriate; (4) a feeling that school-level criticisms were not heard; (5) the state was unclear about guidelines for remediation and/or seemed wishy-washy on several issues, resulting in local confusion about the direction and operation of the program; and (6) the belief that the tests were more a political expediency than an educational tool.
- Obviously there were numerous issues that were clearly state-specific. For example, State A districts liked the "no strings" money from the state; State B systems devoted considerable attention to documentation to protect themselves against "probable" lawsuits; State A educators and students took the test

less seriously than State B. Community interest in test results seemed higher in State B; and teaching to the test was viewed as potentially a problem in State A whereas State B educators were confronting the reality of this practice.

The authors used the actual field notes to review the terminology local educators used in discussing the tests. Using the conceptual framework, themes, data summary chart information, and this review of responses, individual questionnaire items were constructed. The items fell into five categories: local internal and external operating contexts, the administration of the tests in the local setting, the strategies used to maximize student performance, the purposes the tests were used for in the local setting, and the impact of the tests on instruction, organization, and culture. Items for the questionnaire were generated. This was the product of Phase 1 data analysis.

Phase 2: Survey Design

The second phase of the study involved a quantitative assessment of the local ramifications of mandatory statewide testing programs. Four major activities were conducted during this phase:

- instrumentation
- sampling
- data collection/processing
- analysis

These are summarized in Figure 3 and are elaborated in the text that follows.

Instrumentation. The objective of this activity was to create a questionnaire that could be self-administered in 20 to 30 minutes and returned to RBS for processing. The work on this activity began in May 1986. In some cases this involved the development of discrete items while in other cases a battery of items were assembled to form a scale.

Figure 3

Key Features of Survey Research Plan

Instrumentation

- develop a self-administered survey
- operationalize issues identified above
  - review of existing literature and instruments
  - adapt results from first phase of field work
  - incorporate consultant advice
  - pilot test

Sampling

- approximately 225 districts
  - all districts in State B (N=24)
  - 40% sample in State A (N=200)
- random sample
- respondents from three role groups (one person per role group per district)
  -central office
  -building administrator
  -teacher
- develop followup procedures and replacement sites

Data Collection/Processing

- mail surveys in fall 1987
- implement followup procedures and replacement sites
- data entry and cleaning

Analysis

- use conceptual framework to guide analysis
- three separate analyses (one for each group)
- two kinds of analyses
  -descriptive (average and variation)
  -relational (association among variables)

The next step in the development of the instrument was a pilot test to ensure that the questionnaire was clear, communicated the intent of the project, and could be completed within time constraints. This activity involved: identifying a small group of teachers, principals, and district administrators; administering the survey; and spending adequate debriefing time to ensure that all concerns about the proposed instrument were aired. Changes to the questionnaire were made on the basis of the criticism that was offered. A sample copy of the instrument is in Appendix B.

Sampling. The unit of interest in this research was the school district. State B had 24 districts and State A had 501. All districts were invited to participate in the study.

Three different role groups familiar with the testing program were targeted for each district: central office administrators, principals, and teachers. A separate questionnaire was completed by each role group member. In state B, where there were fewer but larger school districts, three respondents from each role group within the district were asked to complete the survey. Only one person from each role group within the district completed the survey in state A. The participating staff in each system were selected by the superintendent or a designee.

Data Collection/Processing. Invitations to participate in the study were mailed in the late fall of 1986. A district liaison person whose responsibility was to assure timely return of the completed surveys to the researchers was appointed by each district. Data were collected in the late fall 1986 and early winter 1987. In all, 23 of the 24 State B systems returned useable questionnaires with three respondents for each of three role groups (central office, principal, and teacher) for each district. In State A 277 of the 501 districts responded with one respondent for each of the role

groups. In State A, the response rate far exceeded researchers' expectations. The response was fifty-five percent of the population to which generalization was desired rather than fifty-five percent of just a sample of that population. No followup was undertaken. In State B, phone calls to the district liaison person elicited responses from almost the total population. Both response rates provided more than adequate samples for generalization to the entire state (Krejcie & Morgan, 1970). An analysis of the participating and non-participating districts in state A showed no significant differences between the two groups in terms of basic demographic characteristics (e.g. size, wealth, location).

Analysis. The analysis had three foci. The first was to identify educators' responses concerning the adjustments they had made. The second was to examine cross-state differences on each of the six indicators for instructional, organizational, and cultural system adjustments (see Figure 1). The third was to examine within-state district variations for three of the system adjustments. The statistical steps taken for each are discussed below.

Phase 3: Follow-up Fieldwork

In the fall of 1987, field researchers returned to 11 of the original 12 sites visited in Phase 1. (One State B system did not participate). The purposes of these visits were: (1) to trace subsequent developments in the operation of the state testing program; and (2) to obtain assistance in interpreting the results of the survey. Over 80 local educators participated in this activity. The interviews concentrated on the findings contained in the section on within-state district variations. Essentially, the findings were presented to participants and they then reacted to specific numbers, interpretations, and implications. These reactions then were incorporated into the quantitative analysis sections of this report.

Findings Regarding Educators' Reactions to Statewide Tests

The questionnaire asked school district professionals to report their reactions to state-mandated minimum competency tests. There were five categories of items on the instrument: local context, the tests, purposes, strategies, and impacts. This section of the report contains educators' responses to selected individual items from each of these categories. This presentation gives a flavor of how educators felt about their respective states' programs and hints at important differences between the two states as well as important variation within each state.

Frequency distributions for the entire sample, state-to-state comparisons, and role group comparisons (central office-CO; principal-Prin; and Teacher-Tchr) are provided in the tables for the items.

Local Context

To fully understand the effects of implementing state-mandated testing programs, it is important to understand the contexts in which districts operate. Two important contextual features that may help explain local responses to MCT programs include the degree to which the relationship between local educators and the state legislature is a constructive one and the degree to which the state testing program is seen as a helpful tool for improvement.

With respect to the political climate between the state and local districts, it can be argued that where the relationship was more positive, a more positive effect of the MCT would be felt. Survey respondents offered reactions to the statement:

> State legislators are generally supportive of professional educators.

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
def. false	3	3	2	4	12	11	11	14
prob. false	11	12	9	14	20	13	26	20
not true or false	31	28	34	31	21	22	18	24
prob. true	51	53	51	48	45	51	44	39
def. true	4	4	5	4	3	3	2	3

Five response choices were available from definitely false to definitely true. The numbers in each table represent the proportion of respondents making that response choice. Respondents in both states were moderately positive in their assessment of the supportiveness of state legislators. Approximately half felt the statement was probably or definitely true. Respondents were more positive in their view of state legislators in State B than in State A.

A central purpose of MCT was that the results would provide a standard and systematic method for identifying students who were not performing well and may have needed additional instructional assistance. To the degree that the MCT offered a useful tool for identification of students in the two states, it would probably be more useful. Survey respondents were asked to respond to the statement:

> The district has a well-developed method for identifying students with special needs.

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
def. false	0	-	-	0	2	-	3	2
prob. false	1	1	-	3	2	1	-	5
not true or false	2	2	0	5	5	4	6	5
prob. true	25	21	23	31	28	24	22	37
def. true	71	75	77	61	64	70	68	52

Practitioners almost universally (over 90%) reported that they already had a well-developed system for identifying students with special needs. The implication from these findings was that little would be learned from the MCT as a means of identification. This was confirmed by followup interviews where practitioners reported that they already knew which students needed additional attention:

> No one had to tell us who was having problems. Most of the students identified by the MCT had already been identified. Out of 1100 identified for Chapter I additional basic skills instruction, only 8 new students were identified by the MCT.

> We are able to predict MCT failures with our own tests. The MCT indicated 27.6 percent of other students were failing. Our own tests had previously indicated the failure rate was 26.2 percent.

The Tests

It can be argued that when practitioners viewed the MCT results as accurately measuring student attainment, they would probably take the test more seriously and one would be more likely to find beneficial changes in response to those results. Likewise, a more favorable view of the measurement properties of the MCT would be linked to greater impact on students.

The survey asked respondents to indicate the degree of truth to the following statement:

> MCT gives an accurate reading on student attainment.

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
def. false	4	5	3	4	15	18	11	16
prob. false	13	15	10	13	27	21	29	31
not true or false	25	25	27	21	23	22	24	24
prob. true	53	48	55	56	33	36	35	27
def. true	5	6	4	6	2	3	2	2

The responses were varied. Just over one half of the respondents in State A and one third in State B agreed (probably or definitely true) that the MCT was accurate. Nearly a quarter of the sample was not sure whether the MCT was accurate. One in six respondents in State A and one in three in State B disagreed and felt the test may not provide an accurate reading on student attainment.

Equally important, many interviewees reported that the MCT forced them to focus on some curricular content at the expense of other important areas that were being excluded.

> The tests have forced us to rob Peter to pay Paul. For example, we are spending time with students remediating for the citizenship test, but to do so we have pulled them out of an earth science class. Consequently, they are failing earth science.
>
> Senior high teachers resent having to teach sixth, seventh and eighth grade competencies to their students.
>
> An inordinate amount of time has been spent on [MCT] skills when the focus should be on thinking skills.
>
> Too much time has been taken from stuff [curricular content] I used to do.

This feeling was verified by survey results with the statement that:

> Staff feel there is a discrepancy between what they think should be taught and what the tests emphasize.

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
def. false	6	5	5	7	4	-	2	10
prob. false	37	33	35	42	19	22	16	17
not true or false	27	28	28	23	20	16	30	15
prob. true	24	27	24	22	35	37	34	34
def. true	7	8	7	6	22	25	18	24

The difference between State A and State B was striking. Nearly twice the proportion of respondents in State B than in State A reported some degree of tension (probably or definitely true) between what they taught and what the tests emphasized. Almost no one reported a lack of discrepancy (i.e. definitely false).

Purposes of the Test

One important way in which MCT tests might be used is as a benchmark to judge the effectivenss of a school system. While the results initially were meant to be diagnostic and used only to identify students with additional instructional assistance needs, the scores were often used to compare one system with another. Since there were no universally agreed upon criteria upon which school systems were judged, these readily available and comparable results were often used in this manner. This was a matter of great concern to school districts. For example, a State A rural district with a reputation for quality educational programs commented:

> The public perception of the district is based on test scores...We have a reputation as one of the best districts in the [region], but our test scores don't reflect that. There is increased pressure from the board and community. We must not allow the press to make us look bad again....I am going to have to spend time I shouldn't, doing PR work.

In another district, a principal reported that in response to published school-by-school test score results in the local newspaper, two-thirds of a recent PTA meeting was spent on explaining the results. The majority of parents in this school knew about the school rankings. In State A, the districts have not made a big issue of the results. Indeed, one central office administrator suggested that scores were released to the press only when reporters asked for the results repeatedly. Nevertheless, most

interviewees indicated a strong push by parents and the media to highlight the scores. A principal reported that parents had moved their children from one school to another based on nothing but test scores.

Educators were asked to indicate how frequently the MCT results were used:

> To compare district performance with the performance of nearby school districts.

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
never	13	14	14	9	4	6	3	2
almost never	10	8	14	7	3	7	2	-
seldom	20	20	21	17	11	12	13	8
fairly often	23	22	20	28	25	27	25	25
frequently	20	24	14	22	30	22	30	40
very frequently	15	11	18	17	27	25	28	26

The results indicated that three of five respondents in State A and four of five in State B felt that district-to-district comparisons were made at least "fairly often". The greater frequency as reported by survey results in State B quickly disappeared with the introduction of district-to-district MCT comparisons by the Chief State School Officer in State A.

Another way in which MCT results were used is to alter course content. The two different state approaches were clearly seen in the response to the question:

> Teachers have altered the content of their classes.

	STATE 1				STATE 2			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
no change	19	17	17	25	1	-	2	-
minor change	37	41	36	35	7	9	8	5
moderate change	37	35	40	35	39	38	39	41
major change	7	17	7	6	49	46	51	51
total change	1	-	-	-	4	7	-	3

In State B where the stakes were much higher there was a concerted effort to align curricula with the instructional objectives outlined by the test developers. Over half of the respondents in State B indicated that a major or total change had been made in class content as a result of the MCT. On the other hand, only eight percent of the respondents in State A reported a similar degree of change in course content.

Strategies

School systems employed a variety of strategies in adapting to the demands placed on them by the MCT. Two common strategies, both designed to improve test scores, were to review test content just prior to test administration and to appoint an employee to take primary responsiblity for educating the professional staff regarding the tests.

While many practitioners had a negative view of "teaching to the test", survey respondents indicated that reviewing test content prior to the exam happened regularly. This was particularly true in the high stakes setting of State B where students would be denied diplomas if they failed the MCT. Respondents were asked if:

> Content and skills covered on the MCT are reviewed just prior to test administration.

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
def. false	42	43	43	38	-	-	-	-
prob. false	29	29	29	29	3	1	5	3
not true or false	10	10	10	10	5	4	6	3
prob. true	15	15	12	17	38	37	47	31
def. true	5	3	6	6	54	57	42	63

Almost all the respondents in State B indicated that this was probably or definitely true. Considerably less emphasis was placed on this strategy in State A where the MCT had lower stakes.

Another common strategy was the focus of this question:

A person has been put in change of MCT-related staff development activities.

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
def. false	15	17	15	13	3	6	2	2
prob. false	10	8	7	15	4	3	2	7
not true or false	10	9	13	8	12	15	15	6
prob. true	17	15	19	17	19	17	12	31
def. true	48	50	46	47	61	59	70	55

A key indicator of the significance an organization places on an activity is whether it creates a separate job or clearly delineates a person with the responsibility to oversee that activity. In both states at least two thirds of the respondents indicated this statement was probably or definitely true. It was clear from these numbers that local systems were taking the MCT activity seriously and placing some emphasis on the need to address staff development related to the MCT.

Impacts

As described in other sections of this report, the researchers were interested in documenting a number of different changes at the local level. In particular, emphasis was placed on the impact on students, teachers, and curriculum and instruction. A sample item from each area helps illustrate those changes.

First, the impact on students was assessed. To the degree that the MCT has any significant meaning to students one might expect to find:

Students are more serious about their classes.

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
no change	35	37	31	37	18	12	16	26
minor change	32	33	31	32	28	26	24	34
moderate change	30	29	36	25	44	47	55	28
major change	3	1	2	5	11	15	5	12
total change	0	-	0	0	-	-	-	-

The responses to this question indicate that there had been only minimal impact. Nobody responded that there was a "total change" and only a small proportion suggested a major change. Again, as might be expected since the consequences of failure were higher in State B, respondents reported more impact than in State A. It is also interesting to note that building and central office administrators perceived the MCT as having more impact on students than did teachers in State B.

A second area of change focused on the quality of teacher worklife. One of the items assessed the degree to which:

> There is a decreased emphasis on using educators' professional judgment in instruction matters.

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
no change	61	58	64	63	24	16	31	26
minor change	23	28	20	22	24	35	19	16
moderate change	12	11	15	12	30	27	31	34
major change	2	3	1	4	19	19	17	20
total change	0	0	0	-	3	3	3	4

The concern was that MCT programs would narrow the definition of what was important in the curriculum and would greatly reduce teachers' discretion in how to teach students. The responses for this item suggest that a "moderate change" had occurred with decreased emphasis placed on educators' judgments. Again, the finding was stronger in State B where the MCT had been in place for a longer period of time than in State A.

For the third area of impact, curriculum and instruction, two items are highlighted. The first addresses whether the curriculum had improved and the second whether the curriculum had been narrowed. With respect to improvement, respondents in State A indicated only a "minor change" while in State B the modal response was "moderate". (The terms "minor" and "moderate" were on the item scale in the questionnaire.)

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
no change	27	28	26	29	16	21	15	11
minor change	32	34	30	33	29	26	27	33
moderate change	29	28	33	27	43	38	46	46
major change	10	10	10	9	12	14	12	9
total change	1	1	1	2	1	2	-	2

In followup interviews, it was clear that "improved" was interpreted in very specific ways. Some of the more frequent adjectives used by educators in State B, which they felt were better terms than "improved", included:

- structured
- coordinated
- more focused
- more defined
- sequential ordering
- more systematic
- consistency
- created a consciousness (about what was being taught)

All of these referred to a tightening up of curricular content. What was missing was any judgment about whether the system was better off.

With respect to narrowing of the curriculum, there were marked differences in response between educators in the two states. In State A, approximately two thirds of the respondents indicated there was no change with

respect to curriculum narrowing. On the other hand, in State B only one of seven respondents indicated no change; two thirds of them reported a moderate to total change.

	STATE A				STATE B			
	All (N=824)	CO (N=277)	Prin (N=273)	Tchr (N=274)	All (N=195)	CO (N=69)	Prin (N=63)	Tchr (N=63)
no change	68	66	62	78	14	6	22	16
minor change	22	24	25	17	23	26	27	14
moderate change	9	10	12	4	34	36	28	36
major change	1	1	1	-	22	23	17	27
total change	0	-	-	1	8	9	7	7

The findings presented above offer a snapshot of local educator's reactions to the initiation of statewide mandatory minimum competency tests. There is still much to be learned. The item level findings hint at important differences between the two states. They also suggest a great deal of district-to-district variation within each state. Each of these two issues is addressed in more detail in analyses presented in the next two sections.

Finding Regarding a Comparison of Testing Programs in Two States

The two states designed their testing programs such that there were several important differences (Table 1). First was State B's making a passing score on all four tests a prerequisite for graduating with a diploma. At the time of this report, the first cohort of students required to pass all four were juniors. Special education students who did not meet this requirement could receive a certificate of attendance. In State A, failing students were identified and supposed to receive remediation to be determined by the district. However, students were not required to retake the test until a passing score was achieved. Second, State B tested students beginning in ninth grade, although a practice test was administered in the middle school. State A students took their tests in the third, fifth and eighth grades.

Under both systems, respondents reported confusion over where the instructional responsibility for improving students' performance resided. In State B, high school staff felt they took the credit or the blame for instruction provided in the earlier grades; in State A, students taking the tests in eighth grade would be remediated by staff in a school different from the middle school where they took the test--since failing had no effect on promotion to high school. Third, State B offered no financial assistance for remediation efforts whereas in State A the state legislature made a special $28 million appropriation for this purpose. Fourth, State B initiated a statewide curriculum improvement program several years prior to beginning the testing program with the expressed purpose of anticipating the instructional quality necessary to perform well on the tests. Moreover, educators from around the state were selected by the SEA to provide input into the content and form of the tests. State A's test was a legislative response to the calls for educational reform that accompanied the reports from the commissions and panels convened in the early 1980's and commercial test publishers were invited to bid on a contract to develop the state's instrument.

Clearly, State B's program should have had a greater impact on its local systems than State A's program, primarily because State B's policy insinuated itself into an important organizational event--graduation--and because preceding statewide improvement and actual test development activities engendered a cummulative anticipation of the day the tests would be put into place. One could argue that the chief organizational task of schools is to move students efficiently from grade to grade to graduation. Obstacles to performing this task smoothly have serious ramifications for the allocation of staff and instructional resources, so serious in fact that innovations that blur evidence of progress--such as non-graded classrooms--have little hope of

Table 1

Summary of Two Mandatory, Minimum Competency, State Testing Programs

STATE	TEST CONTENT	GRADES TESTED	PARTICI-PATION	STATE FOCUS	LOCAL CONSEQUENCES
State A	Reading Math	3, 5, 8	Mandatory	Use of test results to identify students in need of additional instruction	Additional funds for low scoring students
State B	Reading Math Writing Citzenship	7 (Practice) 9 10-12 Retests	Mandatory	Identification of failing students to aid districts in curriculum planning	Students must pass test to graduate; LEAs required to provide appropriate assistance to failing students

implementation (Schlechty, 1976). State B's tests posed an obstacle that could not be ignored. Concern about a growing bottleneck of non-graduating seniors (and the inevitable public outcry) coupled with several years of conversation about the tests' arrival sharply focused local educators' attention.

On the other hand, State A's program derived from a dialogue limited mostly to state level legislators and officials. Information reached local educators after the fact. The limited knowledge about the program plus its lack of implications for school operation seemed to insure that the test would have little impact beyond its stated purpose as a means to help schools identify students in need of additional instruction that may have slipped through the cracks.

Areas of Impact

Survey results bear out the expectation that the impact of State B's testing program would be greater than that of State A's. In the remainder of this section, evidence for this conclusion is presented. Specific areas of impact are described first and, then, comparisons between the two states are made. These comparisons are made using the combined responses of equal proportions of three role groups--central office, principal, and teacher--in each state. The reader should keep in mind that post-survey developments would have altered the responses. These effects are discussed at length at the end of the statistical comparisons.

Six clusters of items related to system adjustments evolved from the conceptual and empirical examination of the individual survey items. The three categories of system adjustments outlined in the conceptual framework each had two clusters of empirical indicators. The instructional adjustments included: (1) focused strategies used to improve test scores and, (2)

curriculum and instruction effects. The organizational adjustments involved: (3) the centrality of the test as an indicator of system performance, and (4) information sharing. The cultural category of system adjustments included: (5) student impacts and (6) quality of teacher worklife.

The "Focused Strategies" cluster provided an estimate of the intensity of a system's instructional effort to improve the test scores. Items in this cluster assessed how true each of these statements were:

- Students take a practice test at some point before they take the actual [state] test.
- Content and skills covered in the [state] test are reviewed just prior to test administration.
- The district has provided assistance (e.g. in staff meetings, in-service sessions, and other activities) to help staff identify ways to improve [state test] scores.
- Staff development resources have been allocated to [state test] related activities.
- Special effort has been put into working with the schools in the district where [state test] scores have been lower.
- The entire district is making an all-out intentional effort to improve its [state test] scores.

"Curriculum and Instruction Effects" included items related to the extent of adjustments made in course content and teaching practices. Four items concerned how often the test was used for the following purposes:

- To identify instructional objectives/content already being addressed in the curriculum that were in need of greater emphasis.
- To identify previously unaddressed instructional objectives/content that need to be added to the curriculum.

- To determine student placement in instructional groups within a class.
- To determine student placement in homogeneously grouped classes or courses.

Four items concerned the magnitude of change:

- Teachers have altered the content of their classes.
- Teachers have adopted new instructional approaches.
- Staff members have been introduced to important new instructional ideas.
- Basic skills instruction has spread throughout the curriculum.

The centrality cluster, labeled the Benchmark Effect, involved the extent to which the test scores were becoming important points of comparison for organizational performance within the school, district, and community. Individual items assessed how often the test was used for the following purposes:

- To compare the performance of individual classrooms within a school.
- To compare the performance of individual schools within the school district.
- To compare district performance with the performance of nearby school districts.
- To publicize the school district's performance to the local community.

"Information flow" captured the extent to which aspects of the testing program were discussed or shared within the organization and with the community. Individual items addressed how true each of the following statements were:

- Parents are aware of when their children will be taking [state test].
- Parents receive information on how well their children performed on [state test].

- Teachers receive information on how well their students performed on [state test] overall.
- [state test] scores are a topic of discussion at staff meetings.

The "Student Impacts" cluster was not intended to be an all encompassing category. But the composite of items included in it offered a glimpse of how the culture of student life fared under the test program in terms of the extent of change in each of the following areas:

- Students are more serious about their classes.
- Special education students are receiving increased, beneficial attention.
- Teachers have more empathy for students who are achieving poorly.
- Staff members know more about students who have serious learning problems.

Similarly, the "Teacher Worklife" category focused on the extent of change in important conditions that define the working culture, such as:

- There is a decreased emphasis on using educators' professional judgment in instructional matters.
- Time demands on staff have increased.
- Staff members have been reassigned.
- Staff members are under pressure to improve student performance.
- Paperwork has increased for staff.
- Staff members are more worried about the potential of a lawsuit.

Once again, this measure is not all inclusive of aspects of work generally discussed in the literature under this heading but at least the items are somewhat indicative of whether teachers' worklives were affected.

Table 2 provides descriptive statistics for the full sample of respondents (N=1017). Three findings are worth discussing in more detail. First,

Table 2: Descriptive Statistics for the Cluster Scores (N=1019)

Cluster	No. of Items	Metric	Mean	SD	Scale Range	Cronbach Alpha
Focused Strategies to Improve Scores	6	Definitely false (1) to Definitely true (5)	3.35	0.91	1.00-5.00	.76
Curriculum and Instruction Effects	8	4 items - Never (0) to Very frequently (5) 4 items - No change (0) to Total change (4)	2.09	0.82	0.00-4.13	.82
Benchmark for Comparative Purposes	4	Never (0) to Very frequently	2.17	1.06	0.00-5.00	.71
Information Flow	6	Definitely false (1) to Definitely true (5)	4.52	0.52	1.83-5.00	.65
Student Impacts	4	No change (0) to Total change (4)	1.05	0.63	0.00-3.00	.72
Teacher Worklife	6	No change (0) to Total change (4)	0.98	0.69	0.00-3.43	.80

there is evidence of reliability for each of the six clusters. The far right column reports the Cronbach alphas (Cronbach, 1951) which is a measure of the internal consistency of the cluster. With the exception of the last cluster (information flow), which is on the low end of the acceptable range, the alphas were consistently high.

The second noteworthy finding concerns the wide variation of responses for each cluster score. When reviewing the range of scores, the span of responses closely resembles the maximum potential range. This wide variation is especially surprising in light of the fact that the cluster score is computed as an average across all the items that make up the cluster. For example, in order to score the minimum of 1.00 on the cluster "Strategies" a respondent would have had to answer definitely false (1) to all six items in the scale.

The final important result in the table focuses on the means or measures of central tendency. In one case, Information Flow, there was a very high score reported by most of the respondents. The average of 4.52 is near the top of the 1 to 5 scale. In three other cases (Strategies, Benchmark effect, and Curriculum and Instruction), the means are closer to the mid-range of the potential responses. The final two variables, which both directly assess the amount of change in the district, have means near the low end of the potential continuum.

To ensure that these six clusters represent independent concepts, the correlation matrix of the six clusters was examined. The results, correlations between .10 and .50, indicate a moderate association but also evidence of independence.

State Comparisons

The results in Table 3 assess the differences between the respondents from the two states in the study. The statistical tool used to assess the differences between the two subgroups was analysis of variance. A statistically significant difference is indicated when variation between the two states is high relative to within state variation. For this analysis, six cluster scores for each respondent were computed. Then an analysis of variance was conducted on each of the six clusters, comparing respondents from the two states.

The findings were striking and consistent. In five of the six clusters, statistically significant differences between the states were found. Essentially, school systems in State B focused more directly on improving their test scores, altered their curriculum to a greater extent, and used the scores more often to compare performance among schools as well as between school systems than their colleagues in State A. In the case of the Strategies employed, State A's mean was at the middle of the five point scale whereas State B's was only a halfpoint below the high end. This indicated a high level of attention to improving the scores in State B in absolute terms as well as in comparison to State A. With respect to Curriculum and Instruction adjustments, in both states the tests primarily were used to identify instructional objectives already in the system in need of greater emphasis. The other curriculum and instruction impacts were more modest, basically the difference between a minor one in State A and a slightly less than moderate one in State B. With respect to comparative uses of the test, the difference in the mean scores of the two states was that between "seldomly" using them for comparison purposes on the whole in State A versus

Table 3: Analysis of Variance Comparison of Cluster Scores by State (N=1019)

Cluster	Mean State A	Mean State B	F	Signi.
Strategies	3.10	4.44	393.4*	.000
Curriculum and Instruction Effects	1.94	2.75	148.7*	.000
Benchmark Effect	2.01	2.89	97.9*	.000
Information Flow	4.52	4.52	0.0*	.927
Student Impact	1.09	1.65	185.7*	.000
Teacher Worklife	0.86	2.12	478.5*	.000

* Indicates significance well beyond the .001 level.

"fairly often" in State B. The comparative use of the scores increased dramatically in both states as attention turned from internal comparisons to comparisons with other systems and to publicizing system performance to the community.

There was only one case, Information Flow through the organization, where respondents from both states reported similar levels of emphasis. The means indicate that a relatively high level of communication about the testing program took place. It is not possible to say, however, whether this level of communication was any higher than it typically was for any other aspects of school operation.

The two cultural clusters concerned Student Life and the Teacher Worklife. In both cases, State B respondents reported a greater impact. These statistically significant differences represent the difference between slightly less than "minor" change in State A to a not quite "moderate" change in State B. (The reader is reminded that these two labels appeared on the questionnaire.) In absolute terms, then, dramatic change did not accompany the onset of the testing programs.

Another way to observe these differences involves a comparison of the proportion of respondents in each state who scored above the total sample mean for each of the six clusters (Table 4). In all but one of these cases, the State B proportions were more than double those of state A. These results further confirm the notion that there were significant differences in the two states.

In addition to the six intermediate effects, the instrument also asked respondents to assess whether the adjustments and changes were for the better. Four important indicators of these effects were measured. The items were:

- the degree to which the curriculum has improved,

- the degree to which the system is more interested in improving overall student learning than increasing a specific set of test scores,
- the degree to which the curriculum has been narrowed
- the degree to which staff feel there is a discrepancy between what should be taught and what the tests emphasize.

Each of these was measured by a single item in the survey.

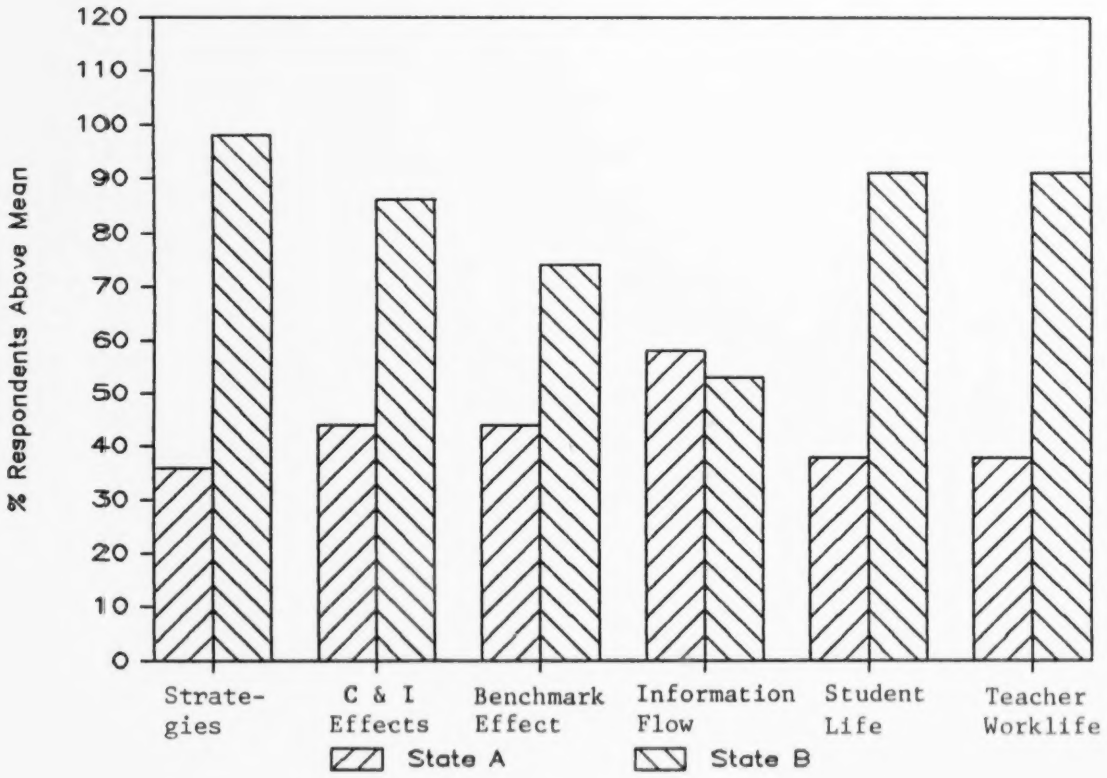
An analysis of variance comparison of state differences summarizes those differences (see Table 5). As with the earlier state comparisons in Table 3, the differences were dramatic and consistent. In State B there was a much stronger feeling that the state mandated testing program had narrowed and yet improved the curriculum, focused the system's attention more on testing than learning, and created a greater sense of discontinuity between the testing program and what staff felt should be taught.

Essentially, the two states had different intentions in mind when the testing programs were initiated and the study data indicate that both were being met. State A wanted to increase the visibility of students who may have been in need of additional instruction and originally had no expressed interest in drastically revamping school programs. State B very consciously wanted to affect the curriculum--first through a planned improvement process and then via the graduation tests. These data reflect the differences in the modest versus the more ambitious approaches.

The dramatic differences between the two states may give the reader the impression that dramatic change occurred in State B. This was not the case. Compared to State A, the differences were great; but if the actual means of the clusters for Curriculum and Instruction, Student Life, and Teacher Worklife are examined relative to the metrics used in the survey, the

Table 4

Proportion of Respondents Scoring Above
Group Mean for Each Cluster, by State



differences in perceived change between States A and B were those between minor change and moderate change (with average scores from one to two on a scale from zero to four). The tests did instigate instructional, organizational, and cultural adjustments at the system level, but these rarely seemed major to the respondents--at least at the time of this study.

Table 5

Analysis of Variance Comparison of Effects by States (N=1019)

Effect	Mean State A	Mean State B	F
curriculum narrowing	0.42	1.83	448.4*
improved curriculum	1.25	1.54	12.2*
focus on student learning	4.36	3.84	61.7*
discrepancy between tests and teaching	2.90	3.52	49.3*

* Indicates significance at or beyond the .001 level.

Recent Developments in the Two States: Raising the Stakes

The above comparisons present a snapshot of the differences in educators' reactions to the testing programs. The picture was taken in the late Fall of 1986 and the early Winter of 1987. Events in both states subsequent to the survey obviously could have had an effect on staff sufficient to alter the responses made on the questionnaire. The feedback visits conducted in Phase 3 of the study elicited comments suggesting the results from State A would be altered dramatically were the questionnaire to be administered in the Fall of 1987. Specifically, the mean scores from State A likely would move closer to those in State B, especially in terms of the Curriculum and Instruction and

Benchmark adjustment categories. State B mean scores also likely would be somewhat higher for both of these categories since all of its systems would be closer to the date when students had to pass all four tests to graduate.

State A. The key event in State A was the publication of the results from the spring of 1987 test administration. Rather than the customary low-key sending of the scores to districts for each to handle as it saw fit, the event was orchestrated by the Chief State School Officer (CSSO). In a public media briefing, the CSSO provided documents that ranked schools in the state from top to bottom in terms of the percentage of students who passed the cut-off point. In addition, a subpopulation of schools that had achieved 100% passing rates despite a "high risk" student population was singled out as being "poised on the brink of excellence." And to cap off the presentation, the CSSO touted the tests as the best measure available to assess the effectiveness of State A's schools. An immediate protest to this use of the scores arose from educators across the state and resulted in the withdrawal of the documents containing the rankings.

The withdrawal of the rankings did not strike the event from either educators' or their communities' emotional record. Educators in three of the six State A districts visited in Phase 3 argued that the "game" had now changed in their systems:

> The purpose of the test changed in September. It is no longer for remediation but to rank order schools. [District 1 superintendent]
>
> The results should be between the state and the school district if the test is to help. When they release scores and say 58 kids need help, we can say we've already identified 40 of them. But the negativism starts; it starts [phone] calls and there is no question I now have pressure on me. [District 2 superintendent]
>
> The test was not all that important....But we might as well face up to it; with the publication of school by school results....one of the goals will be to raise the percentage above the cut score. [District 3 assistant superintendent]

Of the remaining three districts, one--an urban system--had "bought into" the test early in the program and had already begun using the scores comparatively. In fact, interview subjects, to a person, pointed with pride to several of the schools that had achieved "high" passing rates relative to the student population they served. The visibility of the scores was already considerable in the community and the CSSO's actions contributed little additional publicity to how the schools were doing. In another district (which was rural), the community had taken little interest in the scores and, according to the superintendent, the system did not need to treat the test as other than a means of identifying students for additional instruction. In the third, an assistant superintendent claimed that "the publication of scores was deplorable; it was never the intent to rank schools," but that the scores would be downplayed in the district as they had been in the past.

What really seemed to be changing for the first set of three districts in State A was the stakes; they got higher, primarily through the increased visibility of score comparisons and the subsequent increased, albeit reluctant, acceptance of the scores as a benchmark--that is, as a widely recognized point of reference when discussing the performance of schools in the district and surrounding districts. Staff in the three districts reported that they did not believe the tests to be particularly important educationally and did not embrace the tests as valid indicators of attainment. (No staff member in any of the six districts indicated that their assessments of the tests' validity had changed since the survey.) They nevertheless acknowledged that they already were or would soon be treating the scores more seriously than in previous years.

A central office administrator in District 3 commented,

> The tests are not all that important. We use our own standardized testing program to modify instruction.

But since the publicity surrounding the scores increased, more attention had been given to the tests. According to the administrator,

> One thing we did was to say 'here are the objectives on which the test was developed, look at them and see if they are being covered'. This didn't result in change but now that they [SEA] are publicizing the test scores more people who felt they could put the test aside will look at it and say not only have I covered it but do I feel the students will do well. Before I don't think there was as serious a reaction to analyze and interpret the schools' program as there probably is now.

Additional impetus for emphasizing test objectives in this same district came when a six percent difference in the number of students passing occurred between two middle schools in the system. Despite the fact that both had passing rates above 89 percent, the administrator went on to say:

> We couldn't come up with an answer [for the difference] although the lower [scoring] school said they didn't think they should take it seriously. My response is you'd better. We might as well face up to it. One of the goals is going to be to raise the percentage of students above the cut score; so if you're not now emphasizing the test, you'd better. It may not be a legitimate impact, but it is there. The danger is not keeping it in proportion. We need to understand what the tests' place is and that's the danger in how the results are now being emphasized and publicized.

In District 1, a problem arose when surrounding districts' scores matched those of the system, even though the District felt that its carefully and systematically developed curriculum far surpassed the offerings of those around them. The response?

> We don't believe in the tests that strongly but we will be forced to see all material is covered before the tests. We definitely are going to do it. We won't be caught in the newspapers again. [superintendent]

The brunt of not "getting caught" again was to be borne by the reading program--a recently revised, developmental curriculum. The timing of the test administration required shifting the sequence of topics to be covered. An outraged reading coordinator responded,

> You have to alter a curriculum that is already working well and so we can't follow the developmental process. Kids are already growing in a structured program; but it [pressure to change] comes from the board, community, and adverse publicity.

The superintendent empathized with the coordinator,

> I don't have much faith in the tests. I don't want to change the curriculum, and it's not a major revision, but we've got to do better. Still, it's not the right thing to do to anyone. I don't want to over-react but I'm also going to have to spend time on things I shouldn't have to do as well: public relations, testing meetings--just to make the board feel comfortable. It'll never happen again when we see a worse district doing better than us.

The following actions were to be undertaken in a context similar to the first district where standardized tests had long been an integral part of school improvement.

> We feel you can't toy with nationwide standardized tests. That's what we believe in and our performance has been very good. But over the next seven months, we'll be publishing more things about standardized tests and our interpretations of the [state] test scores.

District 2 administrators also indicated a preference not to alter a systematic process for addressing curricula issues. The district took a cyclical approach, working on one content area at a time according to a long-established time frame. No longer. As the superintendent stated,

> We looked at a natural curriculum picture before September but we will address state priorities because our scores were awful. We weren't surprised; the student population we serve is the same as those at the bottom, the big city populations. We will try to raise scores in the third, fifth, and eighth grades. It doesn't mean they'll be smarter.

Another central office administrator detailed the changes more specifically:

> We are building student anxiety, raising their level of concern. We don't want to do that with low esteem kids so we're talking out of both sides of our mouths for our own political needs. Also, changes in math will be addressed in the normal math curriculum cycle next year but this year we'll go ahead and make the changes in 3rd, 5th, and 8th grades. Essentially the [CSSO] just specified the 3, 5, and 8 reading and math curriculum. There is no local option because we have to spend more time on minimal curriculum than enrichment.

Once again, this district had relied on standardized tests in the past to gauge their instructional strengths and weaknesses. The assistant superintendent noted that,

> In the past we've had more of a focus on [standardized tests]. Now the focus has shifted dramatically because we're looking for higher scores in the 3rd, 5th, and 8th grades on the state tests. They'll have more of an impact than the standardized now.

Clearly these three districts were planning expedient strategies to improve the test scores and just as clearly there was resentment to do so and a concern that what they were doing was compromising some sort of standard of good professional practice. Essentially the message they were giving was that the test scores for them were becoming benchmarks for political reasons, namely to appease school boards and communities who had had the opportunity to see their school systems compared to neighboring districts and did not like what they saw. And no matter how district staff had portrayed their performance in the past, part of that protrayal in the future had to include the test scores. Staff, in other words, were beginning to use the tests as a reference for judging local effectiveness. This development reflected obligation more than acceptance. Perhaps most revealing was the ubiquitous "but" in their comments. Woven throughout the above passages were comments like "normally we do that, but now we have to do this." This syntactical form called attention to staff catching themselves in contradictions between what they publicly professed as good and customary practice and what they found themselves actually doing.

How would the above comments modify the survey results presented earlier? No claim can be justified that says these six districts are representative of the more than 270 State A districts that completed the questionnaire. But it is reasonable to suggest that if--in a sample of six--a suburban, rural, and

small, urban district reported that the stakes had been raised enough to obligate them to treat the tests more seriously, then there were others who would have responded similarly and in sufficient quantity to alter the survey results significantly. The single act of publicly ranking schools would seem to have moved State A's low stakes program to at least one with moderate stakes. Thus, the mean scores for some of the adjustment categories would have likely increased.

The quotes cited above indicate that the categories most likely to change would be Curriculum and Instruction and the Benchmark Effect. Emphasizing test objectives and occasionally altering the sequence of content appeared to be the modal way to insure that test content was covered prior to the test. The major impetus for these activities was the increased comparative use of the scores, both within a district and across neighboring districts. No one directly came out to say that staff would "teach to the test" or conduct practice reviews just prior to the test but there did seem to be a reallocating of staff development time to discussing the tests. Consequently, the Strategies category should also increase moderately as should Information Flow--although the already high mean for this latter category leaves little room for an increase. Perhaps the intensity rather than the frequency of communication was where the change would come. The Teacher Worklife mean would rise moderately as the perception grew that the curriculum was being driven by the test rather than by professional judgment, as pressure to improve scores mounted, and as recordkeeping related to keeping track of which objectives got taught to whom before the test increased. Ironically, the category which should be least affected was Student Life. Nothing changed for the students as a result of the publicity surrounding the release of the scores. No obstacle was placed in their path, except the threat of remediation. As one central administrator said,

> There is no impact on students. There would be if something obtrusive was set up for them, but we concentrated on efforts already in place--regular teachers or special reading teachers--[as forms of remediation].

The conclusion is that were the questionnaire to be readministered this year, State A should look more like State B on at least four of the five adjustment categories on which the two states differed significantly.

State B. An additional question is whether events in State B over the last year would have altered the questionnaire response of its local educators. The general answer is that the State B districts seemed to be sharpening the focus of the Strategies to improve scores and seem to be using the scores more and more as Benchmarks, resulting in augmented pressure on teachers to get students to pass. And these increases may have been keeping pace with the likely gains in State A. No single event dramatically heightened the impact of the tests. That situation could change, however, as the time comes nearer when students will have to pass all four of the tests in order to receive a diploma.

In State B, not all four tests were regarded equally. Phase 3 interviews revealed that educators discriminated between the reading and math tests on one hand and the writing and citizenship ones on the other. The reading and math tests, in State B educators' minds, were adequate measures of basic competence in the respective content areas and covered objectives already well-entrenched in the curriculum. The curriculum development aspect of the state initiative began in the late seventies, and these two tests were the first to be developed, trial-tested, and implemented. Actual local curriculum and instruction changes had been in place for seven to nine years in some settings. By 1987, these alterations had become institutionalized, to the point that interview subjects in four of the five districts argued that the

mean score for C&I may have been too low because staff had forgotten that what was now routine was once novel.

> We made sure everything we tested was in the curriculum. But that was done eight or nine years ago. The changes were already made [well before the survey]. [Central Office Administrator]
>
> The mean [for C&I] is skewed. Reading and math have been implemented for a while. [Teacher]
>
> The changes in my area would have occurred well in the past. [Teacher]

The upshot was that the two tests were no longer obtrusive.

> In reading, there probably hasn't been much change; the same in math. The scope and sequence were already complete and the content match was already there. [Principal]
>
> Math and reading teachers probably don't have much of a problem anymore. [Central office administrator]

Such was not the case for the writing and citizenship tests. Both generated considerable controversy. The writing test did so primarily because staff viewed it as demanding a performance level well beyond that necessary to be minimally competent in writing. The citizenship test's controversial aspect centered around its requirement that students memorize information about local, state, and federal governments--information that even the teachers did not possess without special study. Fueling educators' concerns were the facts that students had much more difficulty succeeding on these two tests and that the time when the first cohort of students would have to pass all four tests to receive a diploma was inexorably approaching. For administrators, teachers with responsibilities in certain grades and in certain content areas, and special education teachers, the pressure to achieve passing scores was building and the impact on their work lives was great. The following comments were representative of the opinions expressed in each of the five Phase 3 districts:

District 1: There is an extensive impact on school administrators: scheduling, record-keeping, and realizing that the number of those who failed has become a measure of performance. It's easy to look at [that number]. [Principal].

A central office administrator in the same district summarized:

> The tests are dictating school life in some major areas. We won't see the full impact until the citizenship and writing tests are taken for graduation.

District 2: If you look at all teachers and kids, we're only talking about 25 percent of the staff and kids who are being affected. [Principal]

District 3: We've changed the whole social studies curriculum. We had to expand the 7th and 8th grade American Studies to include more history (to make up for content not being taught later) and now teach government in the last term of 7th and 8th grades which we did not teach at all as a separate entity in the past. And we have structured in key points in the language arts scope and sequence. [Central office administrator]

> Special education teachers cover all four content areas and feel under pressure only because we have students who don't have the potential and yet are in the diploma program. [Teacher]

District 4: I know in English the test is driving the curriculum, at least in 9th grade. [Teacher]

> It depends on who the teacher is and what the teacher teaches. You can't have a bigger impact than on sequence or inserting a new course. We now offer courses not included before and content that changed from 10th to the 9th grades. With government, the impact is overwhelming. [Central office administrator]

> Within social studies, one teacher feels more than the other depending on their assignment. [Central office administrator]

District 5: The impact has been more with citizenship and writing than math and reading. [Central office administrator]

> In my area, there has been a total revision in instructional sequence in contemporary issues to make sure the content is taught before the test. [Central office administrator]

The above comments indicated a "differentiated" impact of implementing the tests. Some parts of the system were affected little while others felt considerable ramifications. Such a situation caused the mean scores presented

above to disguise this important impact of the tests in State B. The point is that some teacher, administrators, and schools were affected by the test in highly significant ways.

The "discomfort" of subgroups of staff involved with the two controversial tests was increasing, according to staff, in contrast to greater comfort with the other two tests. Essentially staff seemed to be focusing more and more on the percentage of students passing the writing and citizenship tests and adopting expedient methods of improving scores. This "concentrated" approach, used by that cadre of staff identified above as being most affected by the tests' implementation, was apparent in all five systems, especially in schools where the scores were lowest.

District 1 staff reported that considerable time was spent in preparation for the tests:

> We are concentrating more on basics. We are now spending from September to November on basic skills rather than on our developmental program. [Reading teacher]

Another person complained that the writing test's importance was getting out of proportion.

> The test has become the judge of the total system. [English teacher]

Schools with low scores seemed to be getting special attention, as indicated in the following comment:

> When the scores are low, it takes me into the school for the names of the kids who failed. There is no stroking in schools where scores have dropped. Everyone is sitting around with bated breath waiting for the test scores. [Central office administrator]

District 2 central office administrators agreed that the tests were assuming greater importance in the system. The scores assumed a constant presence in their work.

> Of course the tests are benchmarks. I always say it's only one indicator but it is the benchmark. It's reality. [Central office administrator]

> The first question we ask is how we did relative to so and so. [Central office administrator]
>
> Today I have 105 seniors who haven't passed. My anxiety is higher. [Central office administrator]

One administrator believed the pressure was greatest on schools with low scores.

> I'm in the middle. I have no pressures at all. I know I'd feel uncomfortable on the bottom. [Principal]

District 3 seemed less consumed by the tests than other systems. Partly because of its small size, the burden of improving test performance fell on only a few shoulders. Moreover, the district had a history of deflecting the impact of state initiatives. Nevertheless, the tests had to be addressed.

> We're bucking the system here. Many districts moved Civics to the ninth and are testing for it in the tenth. We've had a program for a while in the twelfth grade. But it causes problems with no ninth grade civics class; we're interrupting classes to do a review. [Teacher]
>
> I'm right now panically moving toward the test. [Teacher]

District 4 teachers were concerned about the extent to which passing the test was becoming an expediency in the system.

> We realize a kid is taken out of science every other day for citizenship and will fail science to maybe pass the citizenship test. [Building administrator]
>
> We're just getting them to memorize facts until [the test is given]. [Teacher]
>
> I'm not opposed to the idea of testing. But I'm not so sure we haven't gone overboard, the tail is wagging the dog. The original idea was that there were to be certain standards the student would have to meet, but if the student doesn't pass, people will ask what's wrong within the school and teachers. [Teacher]

These very targetted means for getting students to pass were acknowledged as a necessary evil:

> We've had to do things we didn't want to do. [Central office administrator]

Staff in District 5 reported increasingly frequent interactions concerning how students were doing relative to the tests' objectives. They faced heightened awareness of the scores.

> Teachers feel pressured to meet the superintendent's expected pass rate. [Central office administrator]
>
> In administrators' meetings the talk is about where we rank. Parents let you know. You see it in newspapers. [Principal]

The result was the adoption of very focused strategies to teach test objectives in the classrooms.

> Teachers feel jerked around. The test dictates what I will do in the classroom. [Teacher]
>
> If you deviate from the objectives, you feel guilty, especially if kids fail. [Teacher]
>
> We have materials provided by the county as 'quick help.' We were told 'here's how to get kids to pass the test fast.' They were good ideas but specifically on the test. For example, if the area in a rectangle is shaded, you multiply; if not, you add. [Teacher]

And in response to the above stream of comments, a teacher summarized,

> Talk about games and game-playing!

The above comments would suggest that three of the adjustment category means had the potential for noticeable increases: (1) C&I--as the spread of writing instruction throughout the curriculum over the last year affected more than a subgroup of courses and as social studies content sequences were adjusted to insure coverage of course material prior to the test; (2) Strategies--because the nature of the citizenship test encouraged intensive reviews and focused practice just prior to the test administration and because of the relatively high number of students in remediation; and (3) Teacher Worklife--as the pressure on writing and citizenship teachers to improve the passing rates on those tests heightened.

Conclusions. Prior to the State A CSSO's actions in September one could have safely predicted that the stakes in State B would have increased relative to State A over time as the day approached when all four State B tests would "count". Certainly for specific State B administrators and teachers the pressure would intensify beyond any faced by their counterparts in State A. Even though the greater publicity of scores and more frequent inter- and intra- district comparisons in State A makes the prediction of this widening gap somewhat suspect, it may still be upheld. All of the districts in State B must face the day when students will have to pass all four tests to graduate; the stakes remain high for everyone. In State A the stakes were raised for only those systems where publicity surrounding the scores generated widespread criticism of district performance.

Results with Reference to District Comparisons

Policy implementation is strongly influenced by local setting and state context. That is, the interplay of local setting, state context, and policy are more likely to yield variations in implementation than consistency. Such was the case in the two states examined in this study. The previous section documented a consistently strong effect on policy implementation in high stakes situations. This section explores a more subtle and yet equally important issue: How was the impact of the testing program differentially felt within each state? In other words, what were the differences among local districts within each state that influenced the particular implementation adjustments a district made in response to the testing programs?

Explaining variation in three of the six adjustment categories was the major focus of this part of the study: (1) Curriculum and Instruction (C&I); (2) Student Life (SL); and (3) Teacher Worklife (TWL). These three dependent variables were emphasized because they came the closest of the six adjustment categories to tapping the mainstream of the work of schools. The reader should refer to the lists on pages 38-40 containing the items included in each of the measures to keep in mind the very specific meanings that these three general labels have in the following analyses.

While responses were sought from different role groups in the state comparisons, the responses of central office staff only were used in these analyses. The rationale for this decision was based on an analysis of variance which indicated significant variation between role groups, suggesting that scores should not be combined to obtain an overall district score. It was felt that central office administrators were in a better position to be informants at the system level than teachers or building principals.

Also it should be noted that these results reflect the views of informed practitioners at the time the survey was administered. As the discussion in the section on "Recent Developments" clearly indicates, important changes occurred at a later date that may have had an impact on the factors related to these adjustments.

State A District Comparisons

The descriptive statistics presented in Table 6 summarize the variation in adjustments across districts in State A.

Table 6
Descriptive Statistics for System Adjustment Variables in State A (N=277 central office administrators from 277 districts)

Adjustment	Mean	Standard Deviation	Observed Range	Theoretical Range
Curriculum/Instruction	1.94	0.76	0 to 3.63	0 to 4.50
Student Life	1.02	0.67	0 to 3.00	0 to 4.00
Teacher Work Life	.81	0.54	1.17 to 4.00	0 to 4.00

In all three cases, the range of observed responses is more than 70 percent of the potential range. School districts in State A clearly responded differently to the state mandated minimum competency test.

In response to a question concerning the accuracy of the means, local educators who participated in the feedback sessions generally agreed with their accuracy for last year. However, the developments regarding the public ranking of schools and the Chief State School Officer's increased emphasis on the test scores made them think that all three means would be higher if a survey were taken today. Ample evidence supporting this contention was presented in the "Recent Developments" section above.

Using the conceptual framework presented in Figure 1, four categories of variables were selected that might be related to these adjustments:

- internal environment (e.g. percent white, SES, size)
- state environment (i.e. political climate)

- MCT program characteristics (e.g. MCT has adequate procedures)
- other district adjustments (e.g. MCT used as benchmark, testing strategies)

The four categories included a mix of individual survey items and clusters of items. As a first step in the analysis, simple bivariate correlation coefficients were examined to explore the relationship of these variables with the adjustments to Curriculum and Instruction, Student Life, and Teacher Worklife. Only those variables having a statistically significant relationship ($p \leq .05$) were included in the next phase. Thus, different variables are included in the regressions for each of the three adjustment categories.

As a second step, regression equations were calculated using the four categories of variables. The first group of variables entered into the regression equations were internal environment measures, those conceptually furthest removed from system adjustments (i.e. on the far left side of Figure 1). Subsequent equations added one group of variables at a time until all four categories variables were entered.

Curriculum and Instruction (C&I) Adjustments. Table 7 presents the results of the regression estimates for C&I adjustments. When only internal environment variables are included in the regression equation (Column 1), both district size and whether there was perceived to be a well-developed remediation program for students with special needs were related to C&I adjustments. That is, smaller districts and those with stronger remediation programs were more likely to have staff who reported greater C&I adjustments. When additional variables were added to the equation, size was the only variable that continued to be statistically related to C&I adjustments. The analysis indicates that 11 percent of the variation was accounted for by the variables in this internal environment cluster ($R^2=.11$).

Table 7
Standardized Regression Coefficients for Curriculum and Instruction Impacts with Incremental Addition of Independent Variables in State A (N=277)

Independent Variable	(1)	(2)	(3)	(4)
(1) Internal Environment				
• PERCENT WHITE	-.016	-.040	-.077	-.041
• SES	.168*	.101*	.067*	.010*
• SIZE	-.216*	-.196*	-.221*	-.139*
• REMEDIATION ALTERNATIVES	.180*	.177*	-.121	-.079
• HIGH ACHIEVING STUDENTS	-.002	-.105	-.105	-.093
• MCT READING, GRD5	.059	-.048	-.023	-.044
• MCT MATH, GRD5	.107	.080	.060	.146
(2) State Environment				
• POLITICAL CLIMATE		.429*	.282*	.213*
(3) MCT Program Characteristics				
• MCT HAS ADEQUATE PROCEDURES			.034	-.027
• MCT DUPLICATES OTHER TESTS			.054	.045
• MCT ACCURATELY PORTRAYS PERFORMANCE			.092	.136
• DISTRICT PERSON TO COORDINATE MCT			.240*	.099
(4) Other District Adjustments				
• MCT AS COMPARATIVE BENCHMARK				.181*
• TESTING STRATEGIES				.166*
• INFORMATION FLOW				.219*
R^2	.11	.28	.33	.46
R^2 increment (from previous model)		(.17)	(.05)	(.13)

* $p \leq .05$

The proportion of variance explained increased dramatically when the state environment variable (Column 2) was added (R^2 increase from .11 to .28). The healthier district staff perceived the climate between the district and the SEA to be, the greater the magnitude of local C&I adjustments. This strong relationship held up even after the inclusion of all the other variables in the model.

One MCT program characteristic--whether or not a district person had been put in charge of MCT-related staff development activities--was related to C&I adjustments (Column 3). However, that association disappeared when the other district adjustments were included.

In the last step of the regression analysis (Column 4), the results showed that two other district adjustments were related to C&I adjustments. First, where there was a greater acceptance of the test results as an important benchmark of success, local C&I adjustments of greater magnitude were made. Second, where there was a more frequent flow of communication in the district about the state testing program, the magnitude of C&I adjustments was higher . All of the variables in the regression combined account for nearly half of the overall variation in C&I adjustments. (R^2=.46)

Phase 3 interviewees offered important insights about the influence of the Political Climate and Benchmark factors. The six State A districts varied widely in how positively staff members viewed the SEA and the districts' relationship with it. In one district that had made few C&I adjustments of substance, a central office administrator portrayed the situation as follows:

> The community used to hold us accountable. Now we have people in [the state capitol]. Who are they to think they know what our needs are?...The state has become someone we have to beat rather than a partner to work with.

In another district where there was a very high proportion of students doing very well on the MCT, an administrator argued that it was a "pointless exercise" to make C&I changes based on MCT objectives for fear that "a well balanced curriculum could be overbalanced to a minimalist one." The climate had become hostile enough that administrators in at least one district had joined a battle to exempt the district from the MCT.

On the more positive side, while there was no outright admiration expressed for the MCT program, at least one of the six districts adopted the attitude that the MCT could directly help the district. In this system staff at one school had gone so far as to write lyrics to accompany the song "High Hopes" in an effort to motivate students (and staff) to perform well on the tests and to encourage staff to support necessary C&I improvements. Every day for a month before the test, students and staff heard over the loudspeaker the refrain:

> We have worked and studied so long,
> Hope we don't get anything wrong,
> And as you've probably guessed
> On the test
> We'll do our very best
> Cause we have high hopes...

The use of test scores as an important benchmark for comparing schools' and districts' performance was also greeted with varying perspectives in the six districts. On one extreme was an administrator who buried the test results in a bottom desk drawer when they arrived, arguing that the scores created too narrow a definition of what should be taught and how students with learning deficiencies should be remediated. In the middle, teachers and administrators alike shared a concern that the MCT results were being used as "an absolute measure of effectiveness in schools". District administrators were quick to point out the potential negative consequences of public

disclosure of low test scores. However, there was also acknowledgement of the political reality of needing to address the issue. The comment "We will raise test scores", while not stated quite that boldly by everyone, was a refrain in four of the six Phase 3 districts. On the other extreme was the district discussed earlier where two junior high schools with comparable student populations reported slightly different test score results (an 89 percent pass rate versus a 96 percent rate). Although staff members from the lower scoring school explained that they probably took the test less seriously, the community took the difference in scores much more seriously. Enough pressure was created to cause a central office administrator to respond: "They'd [the school] better take it more seriously next time".

In response to the finding that an increased information flow was associated with greater C&I adjustments, interviewees reported that the most useful information was the sharing of test objectives and the process of evaluating the match between those objectives and those already contained in the district curriculum. Where such information was being shared and there was not a great deal of overlap between curriculum and MCT objectives, there was higher probability of substantive adjustments being made in C&I.

Additionally, staff members in districts where substantial C&I adjustments were made were more likely to perceive these changes as having improved the curriculum (r=.527). Phase 3 interviewees claimed that the improvement was in small districts that had had a "textbook" curriculum previously. The test objectives now provided a rationale and structure for the curriculum. The importance of size in the regression model for C&I would seem to support this contention. There was not as strong a relationship between the magnitude of the adjustments and the degree to which district staff considered the curriculum to have been narrowed (r=.192).

Table 8
Standardized Regression Coefficients for Student Life Impacts with Incremental Addition of Independent Variables in State A (N=277)

Independent Variable	(1)	(2)	(3)	(4)
(1) Internal Environment				
• SES	.223*	.118	.112	.058
(2) State Environment				
• POLITICAL CLIMATE		.421*	.341*	.205*
(3) MCT Program Characteristics				
• MCT HAS ADEQUATE PROCEDURES			.015	-.005
• LOCAL INPUT			-.016	-.010
• MCT DOESN'T DUPLICATE EXISTING TESTS			.109	.062
• MCT ACCURATELY PORTRAYS ATTAINMENT			-.022	-.075
• DISTRICT PERSON TO COORDINATE MCT			.155*	.064
(4) Other District Adjustments				
• MCT AS COMPARATIVE BENCHMARK				-.034
• TESTING STRATEGIES				-.006
• C & I				.477*
R^2	.04	.21	.22	.36
R^2 increment (from previous model)		(.17)	(.01)	(.14)

* $p \leq .05$

Student Life (SL) Adjustments. For the SL adjustments measure, the particular items that comprised the scale implied that the greater the score, the higher the relative improvement in student's lives--that is, they took school more seriously, teachers were more sympathetic to students with such learning problems, staff knew more about which students were having problems, and such students were receiving increased and beneficial attention.

One internal environment variable (see Table 8-Column 1) initially related to improvement in student life was the socio-economic status measure (percent of students eligible for free lunch), but that association disappeared as soon as other variables were included in the regression. As was the case for C&I adjustments, the healthier the climate between the district and the SEA (Column 2), the greater the adjustments in SL. This relationship held up even after state program charactertistics and other system adjustment variables were included in the regression. In fact, this one variable alone accounted for 17 percent of the overall variance in SL adjustments.

One state program characteristic showed a moderate association with SL but that relationship also disappeared when additional district adjustment variables were entered into the equation. Finally, the C&I adjustment measure was associated with SL adjustments (Column 4); the greater the adjustments to C&I, the greater the effect on SL.

The majority of interviewees claimed that there was minimal student impact, primarily because other tests had already identified these students as needing remediation. In one urban district with over 1100 students eligible for Chapter I instruction, only 8 new students failed the MCT who had not been identified as "at risk" by other means. In the suburban districts, there were

few failures. Most students easily passed the tests and thus for them the tests were a non-event. The overall impact on students, therefore, was minimal.

Several people hypothesized in the Phase 3 feedback sessions that the MCT had the potential to have a positive impact on students, but that the political climate got in the way. They argued that if the initial intent of the test had been maintained--that is, to help identify students in need of additional instruction, then "the tests have the potential to be valuable." For example, staff from one district reported that they had an excellent reading remediation program in place, and were inclined to develop a comparable math program using the math scores as a stimulus. However, the local relationship with the SEA was so negative, they were hesistant to use the math results as a rationale to help drive the development of the remediation program.

On the positive side, the extra remediation resources the state provided allowed the districts to offer low student-teacher ratios to the pupils who needed assistance the most. These students were getting the additional instruction that enabled them to achieve some success as well as receiving the personal attention that helped build their self-esteem. One principal reported that "kids come in off the playground just to receive remediation instruction." Another positive impact was that the tests forced teachers to examine the objectives embedded in what they were teaching. While most of them reported that they were unwilling to adjust their curriculum just to help students pass the test (i.e. "teaching to the test"), they did report more reflection about what was being taught and this had a positive impact on students.

A minority opinion, voiced by several people, was that the test's shock value was its biggest impact on students. Failure raised the anxiety level of students, particularly eighth graders, and the threat of being pulled from study hall ("that's putting up front to them what the price of failure is") for remediation was enough to motivate them to do better. Staff in one district argued that 50 percent of the initial failures were able to pass a retest exam simply because they had been shocked into the reality of having to face some real consequences for failure (i.e. remediation classes).

Teacher Worklife (TWL) Adjustments. The results in Table 9 indicate that almost none of the variables contributed to an explanation of TWL. Even at an early stage in the analysis without controlling for other variables (the first column), none of the internal environment variables helped explain variation in TWL.

Only one variable revealed a statistically significant relationship. Where adjustments of greater magnitude were made in C&I, the greater the impact on the TWL. That is, if practitioners reported more focusing of instructional objectives, greater altering of course content and increased adoption of new instruction approaches, then they were likely to report greater demands on staff, increased paperwork, decreased emphasis on educator's professional judgment, etc. Because Table 9 is based on central office responses, separate regressions also were computed using only teacher responses. These produced no additional insights as to explanatory factors related to adjustments to TWL. Overall, the proportion of variance accounted for by the variables from the questionnaire was only 10 percent.

When confronted with the findings on TWL, interviewees were not surprised that means were so low but believed that they would be higher now that the stakes had changed (see the section on "Recent Developments"). However, few

Table 9
Standardized Regression Coefficients for Quality of Teacher Work Life Impacts with Incremental Addition of Independent Variables in State A (N=277)

Independent Variable	(1)	(2)	(3)	(4)
(1) Internal Environment				
• PCTWHT	-.124	-.124	-.122	-.120
• SES	.189	.170	.163	.098
• STUDENTS ASPIRE TO COLLEGE	-.046	-.057	-.059	-.049
(2) State Environment				
• POLITICAL CLIMATE		.075	.122	-.018
(3) MCT Program Characteristics				
• MCT HAS ADEQUATE PROCEDURES			-.028	-.078
• MCT DOESN'T DUPLICATE EXISTING TESTS			-.055	-.088
• CONSEQUENCES FOR FAILURE ARE WEAK			-.126	-.087
(4) Other District Adjustments				
• MCT AS COMPARATIVE BENCHMARK				-.014
• TESTING STRATEGIES				.073
• INFORMATION FLOW				-.035*
• C & I				.338*
R^2	.05	.05	.04	.10
R^2 increment (from previous model)		-	-	(.06)

* $p \leq .05$

explanations were proposed for why almost no factors explained the variation in TWL responses. One explanation offered was that there may be wide variation from classroom to classroom within a school. As one teacher commented: "I would prefer to decide on an individual basis for each student what remediation needs to be done." Another plausible explanation offered by a central office administrator was one of self-blame. Practitioners had been taking a great deal of criticism for the inadequacies of the educational system. Every time a new negative finding came out (e.g. poor MCT results) teacher self-esteem went down a notch and with it the perception that the TWL was getting worse. This explanation, then, rested in the induced societal belief that teachers were not doing an adequate job and not in any local contextual factors or adjustments to the MCT.

State B District Comparisons

Table 10 presents the means, standard deviations, observed response ranges, and the potential response ranges for the central office respondents on the three adjustments categories for State B. The higher mean for C&I suggests that the respondents perceived C&I adjustments were of greater magnitude than those in SL or TWL. That is, there was more change in aspects of C&I (e.g., which objectives were being emphasized, course content, and how students were grouped for instruction) than in aspects of SL (e.g., how seriously students were taking school, how sympathetic teachers were to students with special learning needs) or aspects of TWL (e.g., how much pressure teachers felt they were under to improve test scores, how much paperwork for teachers had increased, and how much teachers felt their professional judgment was being deemphasized). Given the potential response range, the difference between C&I impacts and the impacts on SL and TWL was

the difference between "moderate" and "minor" change, according to the metric on the questionnaire. Nevertheless, school systems clearly responded differently in terms of the magnitude of the adjustments they made in each category; in all three cases the range of observed responses was more than two-thirds the potential range.

Table 10
Descriptive Statistics for System Adjustment Variables in State B (N=69 central office administrators from 23 districts)

Impact	Mean	Standard Deviation	Observed Range	Potential Range
Curriculum/Instruction	2.73	0.67	1.13 to 4.13	0 to 4.50
Student Life	1.63	0.66	0.00 to 2.75	0 to 4.00
Teacher Work Life	1.80	0.72	0.33 to 3.33	0 to 4.00

State B staff commenting on the mean scores during the feedback sessions felt the numbers were too low for two reasons. First, as discussed at the end of the "State Comparisons" section, they identified a "differentiated" impact of the tests. That is, certain teachers and most administrators were believed to be highly affected by the implementation of the tests; others were affected very little. So, the means possibly disguised the huge extremes in the impact of the test. Second, staff argued that many of the changes made in reading and math, while extensive in some cases, had been completed years before the survey and that these changes were routinized. Thus, the changes had long since lost their "innovative" feel.

The purpose of the remainder of this section of the report is to explain why these variations in the magnitude of C&I, SL, and TWL adjustments occurred. Because State B had only 23 districts, the use of regression models like those presented in the discussion of State A was precluded. Instead, the analyses uses simple bivariate correlations and tends to identify more significant factors explaining variance in the adjustment categories than was the case in the stepwise regression models.

Curriculum and Instruction (C&I) Adjustments. Table 11 reports the correlations between the four categories of independent variables (i.e. internal environment, state environment, MCT Program Characteristics, and other system adjustments) and the three dependent variables (i.e. Curriculum and Instruction, Student Life, and Teacher Worklife). Local districts reporting C & I adjustments of greater magnitude also indicated that six factors affected the level of impact. Three of the six are scales comprised of several questionnaire items (Political Climate, Benchmark, and Information Flow). The other three factors are single items from the questionnaire. Essentially, the greater the C&I impact:

(1) the more poorly the district's students generally did on standardized tests.

(2) the more favorably staff viewed their working relationship with the state department and the more strongly they felt that the tests were the product of a sincere legislative concern for school improvement (Political Climate).

(3) the more appropriate and timely staff members believed the testing procedures to be.

(4) the more strongly they felt that the tests did not generally duplicate existing testing programs.

(5) the more widely test scores were being used to compare both schools within the district and the district with other districts, (Benchmark).

(6) the more frequent the flow of information about the tests within the district and between district staff and parents of students (Information Flow).

The strongest factor was a district's political climate--that is, the relationship between the district and the SEA. The more positive this relationship, the greater the magnitude of the C&I adjustments. This relationship suggests that substantive rather than perfunctory changes were made in districts that communicated frequently with the SEA and believed the

testing program to be motivated by a concern for school improvement. Even in a high stakes situation, districts making the most significant changes did so not because of the "stick" but because of their positive relationship with the SEA.

This factor points to a more general phenomenon that may be operating in the relationship between school districts and the SEA. An argument can be made that Political Climate and several other significant factors discussed below may tap the degree of respect and trust between LEA and SEA. Thus, overall goodwill between a school district and the SEA may lead to an acceptance of the testing program and a willingness to make C&I adjustments. As one teacher stated, "If you believe in the test, you're more favorable to things that go with it." The "goodwill" theme reappears and will be developed further in the analyses of SL and TWL.

The other factors made intuitive sense. Less change was needed in districts where students traditionally did well on achievement tests, where objectives of the testing program were already embedded in a local testing program, where the state tests were not often a frequent tool of comparison, and where the state tests were not often a major topic of conversation. In other words, the state tests were not top priorities in such systems (as evidenced by the latter two factors), and there was little worry that the tests would become a problem because of the quality of the students and the overlap between the tests and the curriculum.

However, the section on "Recent Developments" in State B presented earlier indicated that the use of the tests as a Benchmark seemed to be influencing C&I changes more over the last year. This occurrence was mentioned in all five of the Phase 3 systems to varying degrees. This development is notable since staff in all five districts felt that they had

done a good job developing the curriculum in years past. It may be that C&I adjustments were becoming a necessity to respond to the politics of comparing schools and districts.

There was a much stronger feeling in State B than State A that the testing program had both improved and yet narrowed the curriculum (see Table 5). During the recent feedback sessions, staff generally viewed the "improved and narrowed" finding as puzzling. Several ventured the guess that the phenomena applied to districts that, prior to the tests, had "textbook" curricula without clearly defined objectives, and thus, "narrowed" meant more structured--and therefore "improved." Commenting on this finding, a principal stated that "where a system didn't have curriculum guides, the state program [including the tests] was the greatest thing since sliced bread." Whereas in systems that had already invested considerable resources in curriculum development, the attitude was more like that reported by a teacher: "The curriculum was already there; it depends on whether a teacher is using it or not--the tests created a consciousness of what should be the focus."

Also, in their comments about the individual tests, State B staff acknowledged that somes aspects of the curriculum had improved but at the expense of other aspects. For example,

> If improved means teaching specific things we know they need to do then we've improved. But there are restrictions on teachers that there didn't used to be. It used to be if a teacher was good then we gave them a lot of latitude in what they do. [Principal]

> We've narrowed math but have improved our overall goal statement. The tests made us look at materials and sequence. [Central Office Administrator]

> In some ways we've narrowed; it depends on the teachers. There is less taught now in some courses than in others, but we have a terrific curriculum and we have resisted watering it down too much. [Teacher]

When you add something, you have to drop something. [Teacher]

For another teacher, "improved" and "narrowed" had more of a connotation that the system had somehow figuratively chosen both the lady and the tiger. Spending the time to improve some students' performance meant less time to work with others.

> I love what the writing test has done for those who pass it. I has established good writing skills. I hate what it has done to kids who failed. It makes them perceive themselves as second-class. So now we have to take time away from those passing to deal with the failures. So what is bad about the test is very bad.

There were several staff members who could not see how "improved" and "narrowed" could co-exist. As one argued,

> I would agree with the narrowing part of it with respect to English and Social Studies--a definite narrowing of focus. I don't see that as an improvement. We're doing a disservice to certain segments of students.

Student Life Adjustments. Four factors were associated significantly with the Student Life measure (Table 11). Three of these were single items from the questionnaire while one was a scale--the C&I adjustments category. The higher the impact on students:

(1) the more strongly staff believed the tests were mandated because of the inability of certain districts to make themselves accountable for student performance.

(2) the more strongly staff felt the tests provided an accurate reading of student achievement.

(3) the greater the perception that local staff had provided input into the test and its development.

(4) the more adjustments made in Curriculum and Instruction.

Districts reporting the most change in Student Life were those where students took school more seriously, where special needs students were

Table 11
Bivariate Pearson Correlation Coefficients*
For Three Adjustment Variables with Various
Independent Variables in State B (N=69 central office
administrators in 23 districts)

Independent Variables	Dependent Variables: Curriculum & Instruction	Student Life	Teacher Worklife
(1) Internal Environment	• High Achieving Students (-.275)*	———	• Proportion of Minority Students (.262)
(2) State Environment	• Political Climate (.346)	• State mandated test because of deficiencies in some districts (.380)	• Political Climate (.405)
(3) MCT Program Characteristics	• MCT had adequate procedures (.283) • Test does not duplicate existing test (.253)	• MCT accurately portrays attainment (.424) • Local input into MCT (.265)	• MCT accurately portrays attainment (.239) • Local input into MCT (.299)
(4) Other System Adjustments	• MCT as comparative benchmark (.284) • Information Flow (.319)	• Curriculum & Instruction Impacts (.375)	• Information Flow (.233)

* All reported correlations have $p \leq .05$.

receiving increased and beneficial attention, and where teachers were more sympathetic to these students as a result of implementing the testing program.

The first three of the factors related to SL adjustments provide additional support for the "goodwill" phenomenon. Central office staff in districts making more SL adjustments indicated that the "blame" for the testing program fell not on the SEA but on recalcitrant districts. They also had provided the SEA assistance in developing the tests, and--not surprisingly--they believed the tests to be valid. Once again, staff members from such districts seemed to have "bought into" the testing program and believed it had positive effects on the students.

Curriculum and Instruction adjustments likely had an effect specifically on the seriousness with which students regarded school. New and/or reemphasized objectives and content would continually be accompanied with the message that these items had to be mastered to finish school. The post-secondary importance of the information or skills was minimal compared to its utility to help students exit school. Staff reported that student seriousness about school was on the increase and was driven home by the presence of remediation. This presented a rare instance when students were told, "You will learn" as opposed to "You should learn."

Nevertheless, the mean scores in Table 10 show that students were least affected by the implementation of the tests. Of course, few students had yet been denied a diploma because of their failure to pass one of the tests. For the majority of students, the tests were a one-time event. They were given a little review, took the test, and then went on with the rest of school.

For students undergoing remediation, the story was different. A large proportion of these succeeded in passing with minimal remediation. But, for a

small percentage of students, there were repeated administrations of the test, frequent pull-outs from other classes, and constant individual attention. This last item actually had some beneficial effects since it countered the apparent drudgery of remediation. Administrators reported that they knew the name of every student who had failed a test, and students in remediation had much less competition for teachers' attention. Staff in two systems told stories of students who regretted eventually passing a particular test because it meant that the student had to leave what had become a comfortable environment. Essentially, then, these students felt that they had received more attention and sympathy and perhaps believed that school was worth taking seriously for once. In sum, the testing program had most impact on students undergoing remediation.

Teacher Worklife Adjustments. Five factors were related to the extent that teachers' lives had been affected by the tests. Two of these were scales, Political Climate and Information Flow, and the other three were single items on the questionnaire. The greater the impact on teachers:

(1) the higher the percentage of minority students in the district.

(2) the more favorably staff viewed their working relationship with the state department and the more strongly they felt that the tests were the product of a sincere legislative concern for school improvement (Political Climate).

(3) the more strongly staff felt that the tests provided on accurate reading of student achievement.

(4) the greater the perception that local educators had provided input into test development.

(5) the more frequent the flow of information about the tests within the district and between district staff and parents of students (Information Flow).

As pointed out by Serow and Davies (1982), minority students often need special assistance to perform well on MCTs. Thus, teachers serving a higher proportion of these students would likely have more pressure on them to get students to a passing level. The teachers interviewed reported that they felt obligated to see that students had encountered all of the material to be covered on the tests. Such an obligation might also increase the amount of paperwork for teachers, especially in those districts that had established record-keeping systems for students with a risk of failing the tests.

The next three factors once again reflect the idea of "goodwill" toward the SEA and the tests. This finding seems somewhat counter-intuitive. In districts where a positive climate existed between the LEA and SEA, where staff viewed the tests as valid, and where local staff had input into the development of the tests, staff felt under greater pressure to improve test scores, busier, less able to exercise professional judgment, and under a greater threat of a lawsuit for failure to get students to pass. In other words, the more positive the view of the SEA and the tests, the more "discomfort" (using a word suggested by an SEA official) teachers felt.

This finding made perfect sense to staff members in every district. Two interpretations for it were advanced. The first was that the "positive" variables reflected a sense of "ownership" of the testing program and that teacher "discomfort" was the product of hard work to make the program succeed. Two central office administrators from the same district clarified the idea of "ownership." One stated, "It's an ownership kind of thing; you want to see it (the program) work right." Another countered:

> No social studies teacher feels ownership. [Pauses]. But we have had more of a dialogue on the citizenship test than any other Curriculum now affects graduation. Maybe we do feel ownership but more of the problem than of the test.

The second interpretation argued that "good" teachers put pressure on themselves to see to it that students would pass the tests and that this pressure was even greater under the conditions of a favorable attitude toward the SEA and its testing program. As one central office administrator said,

> "The more knowledge about the tests that conscientious teachers got, the more worried they become, the more concerned, and the more self-imposed pressure they felt."

The last factor suggests that the more frequent the message, the more likely the message was taken seriously. Where the message was that the test was important, teachers were likely to put pressure on themselves to insure that students would succeed.

From an objective viewpoint, teacher worklife was considerably worse with the advent of the testing program. However, these same teachers worked in situations where the tests were viewed positively and where students were the most needy. For the program to succeed, those who wanted it to succeed most had to endure more stress and put forth more effort. From the subjective viewpoint of staff members, this behavior was regarded as a professionally sound and appropriate response to the implementation of the tests.

Conclusions

Figure 4 is a revised version of the conceptual framework presented earlier as Figure 1. This revised framework summarizes the major findings of the study and includes issues addressed in the text. Specific variables are drawn from the "State Comparison" and "District Comparisons" sections of the report.

Conceptually, the assumed flow of influence among the variables in Figure 4 is from left to right. That is, factors in the larger system environment such as district context variables and state testing program variables influence how the testing program is perceived in the district. These perceptions, in turn, have an influence on how the district adjusts its instructional program, organizational behavior, and cultural environment. These adjustments may result in changes in the system's effectiveness through improved curriculum, etc. Recent developments in both states have shown, that the framework needs to suggest a reverse flow of influences as well. That is, changes in the system's effectivenss may result in new system adjustments and these changes can affect the system's testing program and the larger environment.

Several important summary points can be made from the framework. First, the study demonstrates the strength of the high stakes/low stakes distinction between the two states. A state program had the greatest impact when the scores, or passing rates, were a critical ingredient in making important decisions, in line with Madaus' (in press) original argument. In State B, the important decision was graduation. However, in State A, public comparisons of the scores of schools also increased the stakes by calling community attention to variations in school performance within and across districts. This single event in State A moved a low stakes program to one with at least moderate stakes.

An important question is: Was this for the better? The qualitative data from Phase 3 of the study suggested that as the stakes intensified in both states, there was a point at which district strategies focused on improving test scores took on the flavor of a single-minded devotion to specific, almost "game-like" ways to increase the test scores. State A districts, in

Figure 4

REVISED CONCEPTUAL FRAMEWORK

SYSTEM ENVIRONMENT

1. District Context
   - Student Population [B]*
     - % minority students
     - low achieveing students
   - Political climate between LGA & SEA
   - District Size [A]
2. State Testing Program
   - Stakes: High/Low
     - tied to graduation [B]
     - public score comparisons [A]
   - Local Input into Design [B]

SYSTEM TESTING PROGRAM

1. Existing Program
   - MCT Duplication
2. Perceptions of Test
   - test validity
   - appropriateness of procedures

SYSTEM ADJUSTMENTS

1. Instructional
   - focused strategies
   - curriculum & instruction
2. Organizational
   - information flow
   - benchmark for performance
3. Cultural
   - teacher work life
   - student life

SYSTEM EFFECTIVENESS (objective measures were not part of the study)

1. Curriculum Improved
2. Curriculum Narrowed
3. Focus on Learning rather than tests
4. Discrepancy between test and teaching content

85

* A letter in brackets denotes the state in which a variable made a significant explanatory contribution. No letter in brackets means the variable made a significant explanatory contribution in both states.

particular, that began to take the tests more seriously reported that they did so for political reasons and not because they believed that they were actually improving their instructional program. Prior to this point, the strategies emphasized more systematic changes in the curriculum. Beyond this point, staff began to respond to questions about effects with the phrase: "Some good things have happened as a result of the tests, but..." Staff members' reservations about the practices they were engaging in to improve the scores followed the "but." This analysis suggests that a high stakes strategy seems to have desirable consequences as long as districts are not put under too much pressure. When the pressure to succeed becomes too intense, a turning point is reached and the positive affects become overwhelmed by negative consequences. The exact turning point would vary from district to district; but it was clear that the test scores were beginning to govern activity more directly, as Minzberg (1983) predicted could be the case when an organizational outcome increases in importance.

That it was the difference in stakes that explained the differences in mean scores between the two states rather than simply the length of time that the state programs had been in place is supported in two ways. One, all indication were that the State A means would have risen with the commensurate increase in stakes; and two, State B informants suggested that time likely had reduced the reported means because educators had forgotten that current routines were once innovations.

Second, the perceived political climate between the district and the state department played a relatively strong role in both states in explaining district variations in the impact of the tests on curriculum and instruction, students, and teachers. Essentially the better the communication between an

LEA and SEA and the more the LEA believed SEA actions were not politically motivated, the more likely it was that the district would: match local objectives to those on the test, alter course content, provide increased and appropriate attention to students with learning needs, and report that teachers felt greater pressure to improve test scores. One interpretation of this finding is that this is a "goodwill" factor which is also closely related to positive district responses concerning the tests' validity and the appropriateness of the testing procedures. That is, some districts, for whatever reason, were favorably disposed toward the testing program, and this general "good" feeling about the program engendered a willingness to make considerable adjustments in local operations. This suggests that the historical relationship between an LEA and SEA may outweigh the particular sanctions built into specific policies, even under high stakes conditions.

Third, two demographic characteristics played surprisingly weak roles in explaining district variations. Socio-economic status (percentage of students on free lunch) of the clientele the district served (urban, suburban, or rural) and the type of community served contributed little to the explanatory power of the regression models in State A and did not appear significant in the bivariate correlations for State B.

Fourth, demographic characteristics were not totally unimportant, however. Noteworthy was the negative and significant relationship between district size and Curriculum and Instruction adjustments in State A. Smaller districts made more C&I changes on objectives and content than larger ones. One explanation offered in feedback interviews suggested that small districts may have relied on a "textbook" curriculum in the past where the instructional program was determined solely by the texts adopted. Subsequent to the state

MCT program such districts had to engage in local curriculum development to better match instruction with test content. Another important demographic variable was the moderate contribution of student population characteristics in State B. This made sense under the "high stakes" condition. Teachers and the instructional program had to compensate for a student population that traditionally had performed poorly in standardized testing situations.

Fifth, the findings demonstrated the need to insert a "System Testing Program" category into the framework. There was considerable district-to-district variation in how much the state MCT program duplicated local testing programs, how accurately local staff believed the state MCT portrayed attainment, and how adequate local staff believed the testing procedures to be. Factors related to these areas contributed to explaining local variation in all three of the Systems Adjustments categories used as dependent variables in the "District Comparisons" section. This finding highlights the adaptability of individual districts in terms of putting programs into place. Systems interpreted the state program differently, a fact of life beyond SEA control. These interpretations affected local perceptions of the need, validity, and "burden" of the state program, which in turn influenced the magnitude of adjustments made.

Finally, the findings show the high significance of the original System Adjustments categories (See Figure 4) in explaining district variation in adjustments in Curriculum and Instruction, Student Life, and Teacher Worklife. Several internal and external environment variables that were significant factors in early steps in the regression analysis for State A became insignificant when the adjustment categories were added. This supports the idea that district response was not predetermined by its demographic

characteristics. Rather, how the testing program was interpreted and implemented locally had the greatest influence on how substantial the curriculum, students, and teachers were affected.

In general, some positive results attended the state testing programs. Educators in both states felt their curriculum offerings had become more defined; they welcomed the additional information on students; and they believed students' skills in some areas were improving. But they had misgivings as well. These concerns all centered around the use of test scores as benchmarks for comparisons among schools and as key measures of system effectiveness. Concerns over the validity of the tests and curriculum narrowing might have been downplayed except for the fact that student performance on the tests was becoming increasingly important in both states. "Getting the scores up" seemed to turn minor concerns into major confrontations between sound educational practices and more questionable test-specific ones. This development seems to bear out the concerns of the educators in Darling-Hammond and Wise's (1985) study about MCT becoming an end in itself rather than a means to greater student learning.

REFERENCES

Airasian, P.W. (1987). State mandated testing and educational reform: Context and consequences. American Journal of Education, 95(3), 393-412.

Apple, M. (1982). Education and power. Boston: Routledge & Kegan Paul.

Berman, P.E. (1981). Educational change: An implementation paradigm. In R. Lehming and M. Kane (Eds.), Improving schools: Using what we know. Beverly Hills, CA: Sage.

Corbett, H.D., Dawson, J.A., & Firestone, W.A. (1984). School context and school change. New York: Teachers College Press.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Darling-Hammond, L., & Wise, A.E. (1985). Beyond standardization: State standards and school improvement. Elementary School Journal, 85(3), 315-335.

Elmore, R.F. (1980). Backward mapping: Implementation research and policy decisions. Political Science Quarterly, 94(4), 601-616.

Gordon, D. (1984). The myths of school self-renewal. New York: Teachers College Press.

Krejcie, R.V., & Morgan, D.W. (1970). Determining sample size for research activities. Educational and Psychological Measurement, 30, 607-610.

Lortie, D.C. (1975). Schoolteacher. Chicago: University of Chicago Press.

Madaus, G.F. (in press). Testing and curriculum: From compliant servant to dictatorial master. Chestnut Hill, MA: Boston College.

Marshall, J.C. (1987). State initiatives in minimum competency testing for students. Policy Issue Series No. 3. Bloomington, IN: Consortium on Educational Policy Studies.

Miles, M.B., & Huberman, A.M. (1984) Qualitative data analysis. Beverly Hills, CA: Sage.

Mintzberg, H. (1983). Structure in fives: Designing effective organizations. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W.J., Cruse, K.L., Rankin, S.C., Sandifer, P.W., & Williams, P.L. (1985). Measurement-driven instruction: It's on the road. Phi Delta Kappan, 66(9), 628-634.

Rosenholtz, S.J. (1987). Education reform strategies: Will they increase teacher commitment? American Journal of Education, 95(4), 534-562.

Rossman, G.B., Corbett, H.D., & Firestone, W.A. (1985). Professional cultures, improvement efforts and effectiveness: Findings from a study of three high schools. Philadelphia, PA: Research for Better Schools.

Rossman, G.B., Corbett, H.D., & Firestone, W.A. (forthcoming). Culture, change, and effectiveness. Albany, NY: State University of New York Press.

Sarason, S.B. (1971). The culture of the school and the problem of change. Boston: Allyn & Bacon.

Schlechty, P.C. (1976). Teaching and social behavior. Boston: Allyn & Bacon.

Serow, R.C., & Davies, J.J. (1982). Resources and outcomes of minimum competency tests as measures of equality and educational opportunity. American Educational Research Journal, 19(4), 529-539.

Shannon, P. (1986). Teachers' and administrators' thoughts on changes in reading instruction within a merit pay program based on test scores. Reading Research Quarterly, 21(1), 20-35.

Stake, R. E., Bettridge, J., Metzer, D., & Switzer, D. (1987). Review of literature on effects of achievement testing. Champaign, IL: Center for Instructional Research and Curriculum Evaluation.

APPENDIX A

Sample Data Summary Charts

APPENDIX A

Data Summary Charts

Researchers during Phase 1 of the study completed a series of data summary charts to facilitate both cross-site comparisons and communication of information among the research team. This appendix contains a sample of each of the charts, with data from a State B district included.

Figure 5

Data Summary Chart

Information Sources By Information Categories

Using the interview guide categories, indicate from whom information has been obtained. Place the number of people talked to within a position category about each topic in the space provided. Also, on the first line give the total number of people within each position category you talked to during the site visit.

Category	Supt.	Other Central Asst Supt.	Other Central Office	Bldg Adm	Teachers	Other Prof. (GCs)	Students	
Total Interviewed (39)	1	1	10	3	8	3	12	1
Local Testing Program				1				
-Planning								
-Implementation								
-Institutionalization								
State Testing Program								
-Levels, Standards	†		††	†	†††	†	†††	
-Competencies							†††††	
-Consequences	†		††††	†	†	††	††† †	†
Internal District Context								
-Instructional -Prior	†				†			
-Effects		†	††††	†††	††††	†	†	†
-Organizational -Prior								
-Effects		†	††††	††	††††	††		
Cultural -Prior	†	†	†		†			†
-Effects								
Environmental Context								
-SEA	†	†	†††	†				
-Community		†		†††	††††	†		†
-Media			†					
Test Scores/Alternative Criteria								

Figure 6

Data Summary Charts

System Adjustments

Briefly describe the adjustments that can be attributed to the testing program. Do this in two parts. First, summarize the particular effects identified by sources without evaluating their merit or using the jargon of the interview guides. Second, identify what you interpret to be the most important effects (which possibly will not totally overlap with those identified by sources) and a brief rationale for selection.

Source-Identified Effects	Sources
See page 2	

Researcher-Identified Effects	Rationale
None other than those identified by sources.	

Teaching to the test	1 Assistant Superintendent 4 Central Office Staff 3 Building Administrators 2 Teachers 1 Student 1 School Board Member
Made us pay attention to what we were doing re content, instruction, curriculum, competencies, achievement	1 Superintendent 7 Central Office Staff 2 Building Administrators
More emphasis on basics/survival skills/functional skills	3 Central Office Staff 1 Guidance Counselor 3 Teachers
Curriculum is now more structured/standardized	1 Superintendent 1 Assistant Superintendent 3 Central Office Staff 2 Teachers
Teachers/administrators are becoming more accountable/ pressured/competitive	2 Central Office Staff 3 Teachers
Reading/writing/basic skills instruction now spread through the curriculum	1 Assistant Superintendent 2 Teachers 1 School Board Member
Overload on guidance counselors' time	2 Central Office Staff 2 Guidance Counselors
Actual or potential increase in dropouts	3 Central Office Staff 1 Teacher
Special arrangements for testing	2 Building Administrators 1 Guidance Counselor
Generation/purchase of instructional materials	2 Central Office Staff 1 Teacher
Upgrades education and learning	2 Central Office Staff 1 Guidance Counselor
More staff development/ greater continuity in	2 Central Office Staff 1 Building Administrator
Inundated by paperwork	2 Central Office Staff 1 Guidance Counselor
Students are more accountable	2 Teachers

Guarantees direction in learning	1 Assistant Superintendent 1 Guidance Counselor
Lack of time to teach other than . ? objectives	2 Teachers
Increased student awareness	1 Central Office Staff 1 Building Administrator
Brought credibility to a diploma	1 Central Office Staff 1 Guidance Counselor
Hard word/extra effort by teachers	1 Superintendent 1 Central Office Staff
Student resentment	1 Central Office Staff 1 Student
Concern about lawsuits	1 Central Office Staff 1 Building Administrator
Replaced an industrial arts teacher with a reading specialist	1 Building Administrator
Makes educators aware of deficiencies in our program	1 Building Administrator
Made us pay attention to student attitudes about their abilities	1 Central Office Staff
Good teachers are getting better	1 Central Office Staff
People have a better feeling about schools now	1 School Board Member
Has done the State Department great good in the eyes of the people	1 Assistant Superintendent
People are pleased about going back to basics	1 Assistant Superintendent
Affected what Chapter 1 emphasizes	1 Central Office Staff
Changed sequence of curriculum	1 Central Office Staff
Caused cohesion among the 24 districts	1 Central Office Staff
Teachers share ideas on how to prepare students for the test	1 Principal
Increased teaching, especially in writing	1 Teacher

Figure 7

Data Summary Chart

Overview of School Context

Make note of data bits obtained on the two context categories listed in the Conceptual Framework. This should really be a "pool of tools" that will be used later in the data analysis.

Internal		Environmental	
Category	Data Bits	Category	Data Bits
Instructional		Community	
	1. Children are encouraged, but not to the extent we would like. Lots of money is put into special education, very little into gifted. We would like to see more emphasis on achievement. (School Board Member)		1. There is excellent parent involvement in the schools -- largest in the state. The school system has encouraged it. Parents didn't feel welcome at one time -- late 1960's and early 1970's. The Superintendent believes in parent involvement. (School Board Member)
	2. There is below average achievement [in the district]. Low college entrance. Conservatives. At one time it was thought not good to have achievers. The staff did not want programs for the gifted and talented. A.P. courses were new only recently. I had trouble getting students released to go to college. This district is the "most against achievement I ever saw." (Superintendent)		
	3. Kids have a "do as little as you can" attitude. (Student)		
	4. The district has been "at the cutting edge of educational TV." (Assistant Superintendent)		
Organizational		SEA	

Cultural

National

1. The nation was ready [for [redacted]]. Research was saying back to basics. (Central Office Staff)
2. Students are more serious, a national trend. They are interested in jobs and careers. The pendulum has swung from "do anything that feels good." These feelings have helped [redacted]. (Central Office Staff)
3. [redacted] was part of the action and the passions of our time. Part of the excellence movement -- the thrust of the mid-1980's. (Assistant Superintendent)

Figure 8

Data Summary Chart

Residual Incidents and Data
Worthy of Note

Some comments by sources, observations by you, and events that occurred will not fit into our original thinking. Some of these may be idiosyncratic to your site and some may be similar to what others have found. Record these below with a brief statement of why you think it is important.

Comment/Observation/Event — Why it is Important

Description of the County

[illegible] District [illegible]y is located in [illegible] State B. [illegible]. With the exception of [illegible] town (population = 113,000), the area is very-small-town/rural. There are 42 schools serving approximately 17,000 students. Achievement is below average, according to the Superintendent. People in the county tend to be life-long residents.

SEA

The State Superintendent has been a leader of [illegible] State Program. He felt so strongly about [illegible] State Program that personnel from the state were assigned to the county. This was the first time this happened in the state. (Central Office Staff)

The State Department has been very sensitive in listening to suggested revisions. (Superintendent)

The State Department staff is not large enough to do both assessment and inservice in writing. Inservice is lagging. (Central Office Staff)

I would have done more groundwork at the State Department to bring together the departments of Assessment and Instruction. They should have brought both departments into [illegible] program. No one would have aligned with one group or another. [illegible] State Program at first was a separate entity. It had priority. It was moved into Instruction. (Central Office Staff)

There was great pressure by the State Board to get things going. (Central Office Staff)

Minds were changed too much at the State Department. (Principal)

The State has sometimes listened more to the testing people than to school staffs (e.g., educators thought the writing test should be given in the middle of the school year, but the testing people wanted it to be administered in the spring). (Assistant Superintendent)

Some people have negative attitudes toward the State because 1) comparative judgments have been made by State about district performance, and 2) communication has sometimes been poor, with feelings being hurt, nasty letters written. (Assistant Superintendent)

I hope that the State gives great help when we get into our first legal battle. (Assistant Superintendent)

The biggest problem has been the Department's change of mind (e.g., decided to test citizenship in grade 10, then moved to grade 9). (Central Office Staff)

There is lack of interaction between the graduation requirements committee and the testing committee. (Central Office Staff)

Program was given the best State Department support ever -- people and money. (Central Office Staff)

A Social Studies Bylaw was passed when Program came in. It said what had to be taught in social studies. Another example of two parts of the State Department not working together. (Central Office Staff)

The State has facilitators and coordinators who are also "policemen." (Central Office Staff)

Level of the Tests

The reading test is too easy. (Superintendent, Central Office Staff)

What's tested is really low level. Eighth graders should be able to pass math. (Teacher)

The tests are really easy. (6 Students)

We can't let Program competencies become all we teach. Program competencies are minimal. We can't get comfortable with success in Program. (Principal)

Functional really means functional. (Student)

Program is almost insulting it's so easy. As much as we teach, it's insulting. (Teacher)

Writing Test

Many, many comments available upon request.

Citizenship Test

The citizenship test is fine, but why make a big deal of this. (Student)

Citizenship shouldn't be tested. Older people wouldn't know the answers. (Student)

The citizenship test has too much fact recall and some things aren't really necessary. (Central Office Staff)

I don't think citizenship is that important. (Student)

Most testing falls in ninth and tenth grades. It seems a logical time to test, but those years have quite a bit of turmoil as far as personal lives. It is almost too much for them especially now that citizenship has been added. It is just facts. Students feel frustrated. (Teacher)

Special Education Students

There is a potential problem regarding special education students. They must pass the tests to get a diploma. (Central Office Staff)

There is a new requirement that special ed students in levels 4, 5, and 6 must be tested. Some in levels 5 and 6 are in homes or emotionally disturbed. None has passed. Testing of them is a "real nuisance." (Central Office Staff)

Miscellaneous

If citizenship is a problem, too, this could be a school of Appropriate Assistance. (Teacher)

Program has been almost a revolution in the state. (Central Office Staff)

The press hasn't been too helpful in the writing test controversy. It reports failures instead of success. (Central Office Staff)

Program shouldn't be a requirement. It should be an "indicator." (Student)

In seventh grade, some things are tested that aren't taught yet. (Principal, Teacher)

Even though it seems we had time [to implement Program], we didn't. (Principal)

Program has caused us to violate something we always believed in [i.e., not teaching to the test]. (Assistant Superintendent)

APPENDIX B

Sample Survey Questionnaire

APPENDIX B

Sample Survey Questionnaire

This appendix contains a copy of the questionnaire. The same items were asked of all three role groups in both states, with the exception of a couple of job-related questions at the end. This sample was the one used with central office staff. The acronym for the testing program has been removed from the sample.

A QUESTIONNAIRE ABOUT TEST: USES, STRATEGIES, AND IMPACTS

CENTRAL OFFICE

Research for Better Schools (RBS), the regional educational laboratory for the mid-Atlantic states, is conducting a survey of statewide, mandatory minimum competency testing programs. As part of that survey, we are asking the cooperation of teachers, building principals, and district administrators to complete this questionnaire. Our interest is in the effects these testing programs have on local school systems. This questionnaire is one phase of a larger study being conducted under a U.S. Department of Education contract. We appreciate your taking the time to answer the items on the questionnaire. The identities of respondents, their schools, and their districts will be kept confidential.

RBS

444 North Third Street, Philadelphia, PA 19123

I. *School districts face a variety of internal and external pressures. Listed below is a series of statements that may describe the context in which your school system operates. Please circle the number that best describes how true each statement is for your district.*

CONTEXT	Definitely False	Probably False	Neither True nor False	Probably True	Definitely True	Don't Know
1. This district has a well-developed method for identifying students with special needs.	1	2	3	4	5	9
2. This district has a well-developed remediation program for students with special needs.	1	2	3	4	5	9
3. Students in this district have always done well on standardized tests.	1	2	3	4	5	9
4. Students in this district are tested too frequently.	1	2	3	4	5	9
5. Staff have a strong sense of pride in the work the district is doing.	1	2	3	4	5	9
6. This district places a strong emphasis on students' performance in basic skills.	1	2	3	4	5	9
7. Our district is more interested in improving overall student learning than in increasing a specific set of test scores.	1	2	3	4	5	9
8. The majority of the students in this district aspire to college.	1	2	3	4	5	9
9. Parents are more interested in the performance of their children than in the overall performance of the schools.	1	2	3	4	5	9
10. A major problem is getting the involvement and support of parents whose children perform poorly in school.	1	2	3	4	5	9
11. The community this district serves is keenly interested in public education.	1	2	3	4	5	9
12. The SEA and our school district are in direct communication with each other frequently on a variety of issues.	1	2	3	4	5	9
13. Assessment of student performance is a hot political issue in the state.	1	2	3	4	5	9
14. state legislators are generally supportive of professional educators.	1	2	3	4	5	9

II. *School practitioners vary in their response to state mandated testing programs. Listed below is a series of statements that may describe tests in your district. Please circle the number that best represents how true each statement is for your district.*

	Definitely False	Probably False	Neither True nor False	Probably True	Definitely True	Don't Know
1. Public school educators in this state had significant input into the development of the test.	1	2	3	4	5	9
2. Test duplicates our existing testing program.	1	2	3	4	5	9
3. Test gives an accurate reading on student attainment.	1	2	3	4	5	9
4. Test is administered at the appropriate time of year.	1	2	3	4	5	9
5. Students take the test seriously.	1	2	3	4	5	9
6. Test is administered at the appropriate grade levels.	1	2	3	4	5	9
7. Staff feel there is a discrepancy between what they think should be taught and what the tests emphasize.	1	2	3	4	5	9
8. scores are returned to the district in a timely manner.	1	2	3	4	5	9
9. The format of reports from the state department of education is informative.	1	2	3	4	5	9
10. A variety of remediation alternatives are available to students who fail	1	2	3	4	5	9
11. All students who fail receive some form of remediation.	1	2	3	4	5	9
12. Clearly stated exit criteria exist for determining when a student has completed remediation successfully.	1	2	3	4	5	9
13. The consequences for students who do not pass are too weak.	1	2	3	4	5	9
14. The SEA is a significant source of assistance for handling any test related issues.	1	2	3	4	5	9
15. TEST is the product of a longstanding interest in school improvement by the SEA	1	2	3	4	5	9

	Definitely False	Probably False	Neither True nor False	Probably True	Definitely True	Don't Know
16. Current state financial assistance for remediation encourages a district to have a high number of failures.	1	2	3	4	5	9
17. Test i is a political creation of the state legislature.	1	2	3	4	5	9
18. ~~Test~~ was mandated because the school districts did not do the job of assessing student performance themselves.	1	2	3	4	5	9

III. Listed below is a set of alternative strategies employed by school districts to maximize student performance on the . test. For each of these strategies please circle the number that best characterizes how true it is for your district.

STRATEGY	Definitely False	Probably False	Neither True nor False	Probably True	Definitely True	Don't Know
1. Students take a practice test(s) at some point before they take the actual test.	1	2	3	4	5	9
2. Content and skills covered in the ; test are reviewed just prior to test administration.	1	2	3	4	5	9
3. Every student is told how important it is to take test seriously.	1	2	3	4	5	9
4. Parents are aware of when their children will be taking test;.	1	2	3	4	5	9
5. Parents receive information on how well their children performed on test;.	1	2	3	4	5	9
6. The district has provided assistance (e.g. in staff meetings, in-service sessions, and other activities) to help staff identify ways to improve ; scores.	1	2	3	4	5	9
7. Teachers receive information on how well their students performed on individual items.	1	2	3	4	5	9
8. The district has provided information to staff about test content.	1	2	3	4	5	9
9. Teachers receive information on how well their students performed on test overall.	1	2	3	4	5	9
10. ; scores are a topic of discussion at staff meetings.	1	2	3	4	5	9
11. Staff development resources have been allocated to test;-related activities.	1	2	3	4	5	9
12. A person(s) has been put in charge of test-related staff development activities.	1	2	3	4	5	9
13. Special effort has been put into working with the schools in the district where ; scores have been lower.	1	2	3	4	5	9
14. The entire district is making an all-out intentional effort to improve its scores.	1	2	3	4	5	9

IV. *Districts use ~~tests~~ in a variety of ways. Please circle the number that best describes how often ~~tests~~ is used in your district for each of the following purposes.*

PURPOSE	Never	Almost Never Seldom	Fairly Often Frequently	Very Frequently	Don't Know
1. To compare the performance of individual classrooms within a school.	0	1	2 3	4 5	9
2. To compare the performance of individual schools within the school district.	0	1	2 3	4 5	9
3. To compare district performance with the performance of nearby school districts.	0	1	2 3	4 5	9
4. To publicize the school district's performance to the local community.	0	1	2 3	4 5	9
5. To identify instructional objectives/content already being addressed in the curriculum that were in need of greater emphasis.	0	1	2 3	4 5	9
6. To identify previously unaddressed instructional objectives/content that need to be added to the curriculum.	0	1	2 3	4 5	9
7. To identify students who may need additional help within the regular classroom.	0	1	2 3	4 5	9
8. To identify students in need of additional instructional help outside the classroom who were not already receiving special services.	0	1	2 3	4 5	9
9. To assess student progress toward attaining school district-determined instructional objectives.	0	1	2 3	4 5	9
10. To determine student placement in instructional groups within a class.	0	1	2 3	4 5	9
11. To determine student placement in homogeneously grouped classes or courses.	0	1	2 3	4 5	9

V. *Mandated testing programs have many implications for districts. This list offers a sampling of the ways tests might have had an effect on your district. Please circle the number that best summarizes the magnitude of the impact of test in your district. In making these assessments you will have to draw comparisons between what it was like in the district before tests and now.*

IMPACT	No Change	Minor Change	Moderate Change	Major Change	Total Change	Don't Know
1. More students are attending summer school.	0	1	2	3	4	9
2. Students facing remediation have less choice in terms of classes they can take.	0	1	2	3	4	9
3. Students are more serious about their classes.	0	1	2	3	4	9
4. Special education students are receiving increased, beneficial attention.	0	1	2	3	4	9
5. Teachers have altered the content of their classes.	0	1	2	3	4	9
6. Teachers have adopted new instructional approaches.	0	1	2	3	4	9
7. There is a decreased emphasis on using educators' professional judgment in instructional matters.	0	1	2	3	4	9
8. Time demands on staff have increased.	0	1	2	3	4	9
9. Staff members have been reassigned.	0	1	2	3	4	9
10. Staff members are under pressure to improve student performance.	0	1	2	3	4	9
11. Paperwork has increased for staff.	0	1	2	3	4	9
12. Staff members have been introduced to important new instructional ideas.	0	1	2	3	4	9
13. Teachers have more empathy for students who are achieving poorly.	0	1	2	3	4	9
14. Staff members know more about students who have serious learning problems.	0	1	2	3	4	9

	No Change	Minor Change	Moderate Change	Major Change	Total Change	Don't Know
15. Staff members are more worried about the potential of a lawsuit.	0	1	2	3	4	9
16. Basic skills instruction has spread throughout the curriculum.	0	1	2	3	4	9
17. The curriculum has been narrowed.	0	1	2	3	4	9
18. The curriculum has improved.	0	1	2	3	4	9
19. There is more money available for needed programs.	0	1	2	3	4	9
20. Average class size has dropped.	0	1	2	3	4	9
21. The district's own testing program has been revamped.	0	1	2	3	4	9

VII. During the current school year, what category best describes your job?

_____ 1. Superintendent

_____ 2. Associate or Assistant Superintendent

_____ 3. Director

_____ 4. Supervisor

_____ 5. Other (please specify: _____)

VIII. Are you?

_____ 1. Female

_____ 2. Male

IX. How many years of experience, prior to this year, have you had:

_____ 1. As a professional educator

_____ 2. Working in your current position

_____ 3. Working in this district